

What Works Clearinghouse™

Standards Handbook,
Version 4.1

Contents

I.	Introduction	1
II.	Randomized controlled trials and quasi-experimental designs	4
	A. Individual-level assignment	4
	B. Cluster-level assignment	18
	C. Other analytic approaches	30
	1. Propensity score analyses	30
	2. Analyses in which subjects are observed in multiple time periods	30
	3. Analyses with potentially endogenous covariates	33
	4. Analyses with missing data.....	34
	D. Complier average causal effects.....	44
	1. Criteria for whether RCT studies are eligible for review under CACE standards.....	45
	2. Overview of the process for rating CACE estimates	48
	3. Calculating attrition when rating CACE estimates	49
	4. Procedures for rating CACE estimates when attrition is low	52
	5. Procedures for rating CACE estimates when attrition is high	53
III.	Regression discontinuity designs	56
	A. Assessing whether a study is eligible for review as a regression discontinuity design	56
	B. Possible ratings for studies using regression discontinuity designs	58
	C. Standards for a single regression discontinuity design impact	58
	D. Applying standards to studies that report multiple impact estimates.....	70
	E. Applying standards to studies that involve aggregate or pooled impacts	71
	F. Cluster-assignment regression discontinuity designs.....	73
	G. Reporting requirement for studies with clustered sample.....	76
	H. Reporting requirement for dichotomous outcomes	76
IV.	Single-case design studies	77
V.	Nondesign components	83
	A. Outcome requirements and reporting.....	83
	B. Confounding factors	85
	Appendix A. Assessing bias from imputed outcome data	A-1

Appendix B. Bounding the baseline difference when there are missing or imputed baseline data	B-1
Appendix C. Additional detail for reviews of studies that present complier average causal effects estimates	C-1
References	Ref-1

Tables

Table II.1. Highest differential attrition rate for a sample to maintain low attrition, by overall attrition rate, under “optimistic” and “cautious” assumptions	12
Table II.2. Absolute effect size at baseline	13
Table II.3. Examples of acceptable approaches for satisfying the What Works Clearinghouse statistical adjustment requirement	15
Table II.4. Three categories of joiner risk specified in review protocols	23
Table II.5. Allowable reference samples for calculating individual nonresponse	24
Table II.6. Acceptable approaches for addressing missing baseline or outcome data.....	37
Table II.7. First-stage <i>F</i> statistic thresholds for satisfying the criterion of sufficient instrument strength	53
Table III.1. Regression discontinuity design study ratings	58
Table III.2. Satisfying the integrity of the forcing variable standard (standard 1)	60
Table III.3. Satisfying the attrition standard (standard 2)	62
Table III.4. Satisfying the continuity of the relationship between the outcome and the forcing variable standard (standard 3)	64
Table III.5. Satisfying the functional form and bandwidth standard (standard 4)	67
Table III.6. Satisfying the fuzzy regression discontinuity design standard (standard 5).....	70

Figures

Figure I.1. Steps of the What Works Clearinghouse systematic review process and the <i>What Works Clearinghouse Handbooks</i>	2
Figure II.1. Study ratings for individual-level randomized controlled trials and quasi-experimental designs	5
Figure II.2. Attrition and potential bias	10
Figure II.3. How the What Works Clearinghouse treats sample exclusions in randomized controlled trials	11
Figure II.4. Review process for cluster-level assignment studies.....	21
Figure II.5. Study ratings for randomized controlled trials and quasi-experimental designs with missing outcome or baseline data	35
Figure II.6. Review process for studies that report a complier average causal effect estimate.....	49
Figure IV.1. Study rating determinants for single-case designs	82

I. Introduction

Education decisionmakers require access to the best evidence about the effectiveness of education practices, products, programs, and policies. It can be difficult, time-consuming, and costly for decisionmakers, however, to access and draw conclusions from relevant studies about the effectiveness of these interventions. The What Works Clearinghouse (WWC) addresses the need for credible, succinct information by identifying existing research on education interventions, assessing the quality of this research, and summarizing and disseminating the evidence from studies that meet WWC standards.

The WWC is an initiative of the U.S. Department of Education’s Institute of Education Sciences (IES), which was established under the Education Sciences Reform Act of 2002. It is an important part of IES’s strategy to use rigorous and relevant research, evaluation, and statistics to improve our nation’s education system. The mission of the WWC is to be a **central and trusted source of scientific evidence for what works in education**. The WWC examines research about interventions that focus on improving educationally relevant outcomes, including those for students and educators.

The WWC systematic review process is the basis of many of its products, enabling the WWC to use consistent, objective, and transparent standards and procedures in its reviews, while also ensuring comprehensive coverage of the relevant literature. The WWC systematic review process consists of five steps:

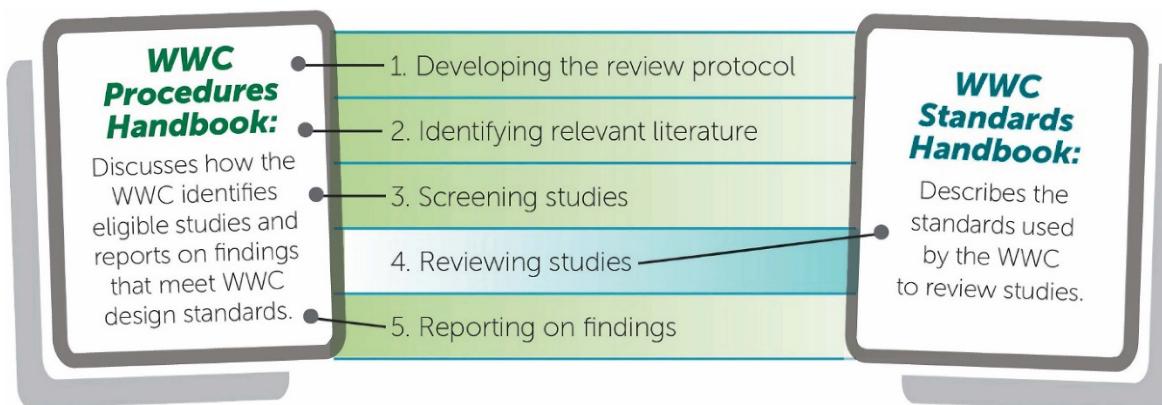
1. *Developing the review protocol.* A formal review protocol is developed for each review effort, including one for each WWC topic area, such as adolescent literacy, primary mathematics, or charter schools. The protocol defines the parameters of the research to be included within the scope of the review, including population characteristics and types of interventions; the literature search terms and databases, if any; and any topic-specific applications of the standards, including acceptable thresholds for sample attrition, risk of joiners in cluster design studies, and specification of characteristics for establishing group equivalence.
2. *Identifying relevant literature.* Studies are gathered through a comprehensive search of published and unpublished publicly available research literature. The search uses electronic databases, outreach efforts, and public submissions.
3. *Screening studies.* Manuscripts are initially screened for eligibility to determine whether they report on original research, provide potentially credible evidence of an intervention’s effectiveness, and fall within the scope of the review protocol.
4. *Reviewing studies.* Every eligible study is reviewed against WWC standards. The WWC uses a structured review process to assess the causal validity of findings reported in education effectiveness research. The WWC standards focus on the causal validity within the study sample—that is, *internal validity*—rather than the extent to which the findings might be replicated in other settings—that is, *external validity*.
5. *Reporting on findings.* The details of the review and its findings are summarized on the WWC website and often in a WWC publication. For many of its products, the WWC

combines findings from individual studies into summary measures of effectiveness, including the magnitude of findings and the extent of evidence.

In addition, the WWC reviews some studies outside of the systematic review process, such as those that receive significant media attention. These reviews are also guided by a review protocol and use the same WWC standards and reporting procedures.

This *What Works Clearinghouse Standards Handbook, Version 4.1*, provides a detailed description of the standards used by the WWC to review studies (step 4). Steps 1–3 and step 5 are described in a separate *What Works Clearinghouse Procedures Handbook*. Taken together, these two documents replace the two documents used since October 2017, the *What Works Clearinghouse Procedures Handbook* and *Standards Handbook, Version 4.0*. Figure I.1 shows how the steps of the WWC systematic review process are divided between the *Standards Handbook* and the *Procedures Handbook*.

Figure I.1. Steps of the What Works Clearinghouse systematic review process and the *What Works Clearinghouse Handbooks*



This *Standards Handbook* provides a detailed description of the standards used by the WWC when reviewing studies that have met eligibility screens, including using one of the following eligible designs: randomized controlled trial (RCT), quasi-experimental design (QED), regression discontinuity design (RDD), and single-case design (SCD). Studies that use other designs are not reviewed by the WWC. The WWC refers to randomized controlled trials and quasi-experimental designs collectively as group design studies. Studies reviewed against WWC standards receive one of the following three study ratings indicating the credibility of evidence from the study: *Meets WWC Design Standards Without Reservations*, *Meets WWC Design Standards With Reservations*, or *Does Not Meet WWC Design Standards*.

The substantive differences between this version of the standards (4.1) and the previous version (4.0) include the following:

- **The “pilot” designation for the SCD standards has been removed and additional clarification around standards elements has been added.** The SCD standards are now approved for use alongside the standards for group designs and regression discontinuity designs. The procedures for synthesizing findings from SCD studies using design-comparable effect sizes can now be found in the *WWC Procedures Handbook, Version*

4.1. The updated standards also include several clarifications, based on WWC [reviewer guidance](#) released in October 2017, that make use of visual analysis techniques to assess whether and how SCD studies meet WWC standards. The SCD standards have also been reordered for clarity.

- **The language of “substantively important” has been removed.** In previous versions of the *Standards Handbook*, an effect size above 0.25 was deemed “substantively important” and noted when characterizing findings. This designation has been removed in this updated version.
- **Small, nonsubstantive changes were made to clarify existing text.**

The remainder of the document is organized as follows. Chapter I provides a general introduction to the *Standards Handbook*. Chapter II provides standards for randomized controlled trials and quasi-experimental designs. This chapter also provides additional standards for randomized controlled trials that present complier average causal effects, with supplemental technical detail in appendix D. Chapter III provides standards for studies that use regression discontinuity designs. Chapter IV illustrates standards for studies that use SCDs. Chapter V provides information on outcome eligibility and confounding factors that applies broadly across designs.

As the WWC uses and applies the standards in this *Standards Handbook*, reviewers may occasionally need additional guidance. If necessary, the WWC will produce guidance documents for reviewers that provide clarification and interpretation of standards and support consistency across reviews. This WWC reviewer guidance will clarify how these standards should be implemented in situations where the current *Standards Handbook* is not sufficiently specific to ensure consistent reviews.

As the WWC continues to refine and develop standards, the *Standards Handbook* will be supplemented or revised to reflect these changes. Any written supplements for use in combination with this *Standards Handbook* will be specified in the protocol governing the corresponding study reviews. Readers who want to provide feedback on the *Standards Handbook*, or the WWC more generally, may contact us at the WWC Help Desk (<https://ies.ed.gov/ncee/wwc/help>).

II. Randomized controlled trials and quasi-experimental designs

This chapter describes the core elements for the review of two major categories of group designs for intervention studies: RCTs and QEDs. While RCTs rely on random assignment to form intervention and comparison groups, QEDs form these groups using methods other than random assignment. Standards are presented separately for studies that assign individuals—such as students—to a condition and studies that assign clusters—such as classrooms or schools—to a condition. The chapter concludes with specific guidance for reviews of studies that use a variety of common analytical approaches.

Although RDDs are sometimes considered a type of group design, the WWC applies separate standards to review eligible RDDs. If a cutoff value on a known measure is used to assign subjects to the intervention and comparison groups, then the study may be eligible to be reviewed as an RDD. The WWC eligibility criteria and standards for reviewing RDDs are described in chapter III.

A. Individual-level assignment

In this section, we describe the three steps for reviewing RCTs and QEDs that assign individual subjects to the intervention or comparison condition:

- **Step 1:** Assess the study design.
- **Step 2:** Assess sample attrition.
- **Step 3:** Assess equivalence of the intervention and comparison groups at baseline, that is, prior to the intervention.

To be eligible for the WWC's highest rating for group design studies, *Meets WWC Group Design Standards Without Reservations*, the study must be an RCT with low levels of sample attrition. A QED or high-attrition RCT is eligible for the rating *Meets WWC Group Design Standards With Reservations* if it satisfies the WWC's baseline equivalence requirement that the analytic intervention and comparison groups appear similar at baseline. A QED or high-attrition RCT that does not satisfy the baseline equivalence requirement receives the rating *Does Not Meet WWC Group Design Standards* (figure II.1). After describing each step in the review process, we conclude with a set of possible results, pointing readers to the appropriate next step in the review process.

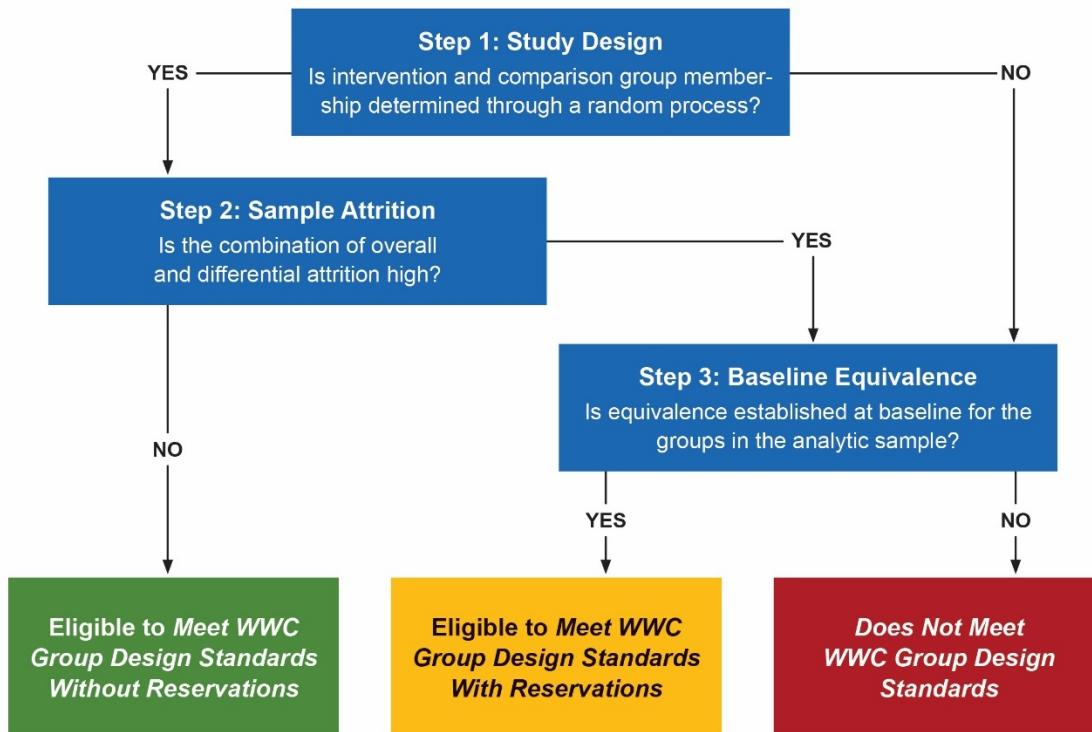
However, individual-level assignment studies that satisfy the requirements outlined in steps 1–3 must also satisfy two additional requirements to be rated *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*. These additional requirements, described in chapter V, are that the study must:

- Examine at least one eligible outcome measure that meets review requirements.
- Be free of confounding factors.

Additionally, when studies use certain analytic approaches, including propensity score analyses, analyses in which subjects are observed in multiple time periods, methods to address missing data, or include endogenous covariates, additional guidance and standards may apply as described in section II.C. In particular, when an analysis uses methods to address missing data

such as regression imputation, maximum likelihood, or nonresponse weights, the review process described in the last subsection of section II.C, *Analyses with missing data*, should be followed instead, which includes an assessment of potential bias from using imputed data instead of actual data. Additionally, standards for reviewing studies that report complier average causal effects are described in section II.D.

Figure II.1. Study ratings for individual-level randomized controlled trials and quasi-experimental designs



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in chapter V, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Step 1. Study design: Is intervention and comparison group membership determined through a random process?

RCTs

The distinguishing characteristic of an RCT is that study subjects are randomly assigned to one of two groups that are differentiated by whether they receive the intervention. Researchers may use any of several possible methods to conduct random assignment. For example, acceptable methods of random assignment include blocking the sample into groups before random assignment, using random subsampling, assigning individuals to groups with different probabilities, and forming groups of different size.

To be valid random assignment, subjects must be assigned entirely by chance and have a nonzero probability of being assigned to each group. Subjects do not need to have an equal chance of being assigned to each group, and the chance of being assigned to a particular group can differ across subjects. However, if subjects are assigned to a group with different

probabilities—that is, if the chance of being assigned to a group differs for subjects within the same assigned condition—then the findings must be based on an analysis that adjusts for the different assignment probabilities. This requirement also applies if the probability of assignment to a group varies across blocks in a stratified random assignment framework.

Compromised RCTs

When the validity of a random assignment process or the analysis of an otherwise well-executed random assignment process is compromised, the study is reviewed using the process for QEDs. There are four ways in which an RCT that assigns individual subjects to the intervention or comparison condition can be compromised.

- The RCT is compromised when the subjects in the analytic sample used to estimate findings were not randomly assigned.
- The RCT is compromised if subjects are randomly assigned to a group with different probabilities, but the findings are based on an analysis that does not account for the different assignment probabilities. Consider a study that conducts random assignment separately within two blocks of students. The study includes the same number of students in both blocks, but students in block A are high performing at baseline, while students in block B are low performing at baseline. The study assigns 70 percent of block A students to the intervention condition but assigns only 30 percent of block B students to the intervention condition. In this case, the intervention group includes 70 percent high-performing students, while the comparison group includes 70 percent low-performing students. If the data are analyzed without accounting for the different assignment probabilities, the dissimilar groups may cause the intervention to appear to have a positive impact, even if it has none. The three WWC-accepted methods of accounting for different assignment probabilities within a group are:
 - Estimating a regression model in which the covariate set includes dummy variables that differentiate subsamples with different assignment probabilities.
 - Estimating impacts separately for subsamples with different assignment probabilities and creating a weighted or unweighted average of the subsample-specific impacts.
 - Using inverse probability weights, formed using the known probabilities of assignment for each subject, as weights in the analysis.

If study authors describe a random assignment process that suggests varying probabilities of assignment but do not make one of these adjustments, then the RCT is compromised and the study is reviewed using the process for QEDs.

- The RCT is compromised when the investigator changes a subject's group membership after random assignment. Consider a study in which some subjects *assigned* to the intervention condition *did not receive* the intervention but remained in the study. For example, some students initially assigned to a classroom implementing the intervention condition may actually attend a different classroom that implemented the comparison condition. If the study authors analyze these subjects as members of the comparison group, based on not receiving the intervention, then random assignment is compromised. However, if the study authors analyze these subjects as members of the intervention group, based on their original assignment—an intent-to-treat (ITT) analysis—then the integrity of random assignment would be maintained. Put another way, not all subjects must actually receive their assigned condition, but all subjects must be analyzed

according to the subject's originally assigned condition. Note that studies that address noncompliance by reporting complier average causal effects (CACE) may be eligible for review using the standards described in section II.D.

- The RCT is compromised when a study author manipulates the analytic sample to exclude certain subjects based on events that occurred after the introduction of the intervention when there is a clear link between group status and the reason for the exclusion. A clear link is present when the exclusion is based on a measure that may have been affected by assignment to the intervention or comparison condition. Not all sample exclusions performed by the author will meet this condition, as illustrated in the following examples. Together, these examples illustrate the three ways in which the WWC treats sample exclusions, summarized in figure II.3: (1) as a compromised RCT, (2) as attrition, or (3) as ignorable, that is, not counted as attrition and not compromising.
 - **Compromised RCT.** If an intervention could affect student attendance—for example, by influencing students' motivation to attend class—and study authors exclude from the analysis students with high levels of absenteeism, then the RCT is compromised. This outcome is represented by the red box in figure II.3.
 - **Attrition.** Suppose study authors grouped students into pairs and randomly assigned one student in each pair to the intervention condition. If either student in the pair was missing outcome data, the exclusion of both students in the pair—or any other larger randomization block—from the analysis would not compromise random assignment because there is no clear link between the intervention and attrition of the pair. In this example, the excluded pair counts as attrition, which does not compromise an RCT and is discussed in detail in step 2 on the next page. This outcome is represented by the yellow box in figure II.3.
 - **Ignorable—not counted as attrition and not compromising.** Some sample exclusions are considered neither attrition nor compromising. For example, if study authors excluded students at random from follow-up data collection, or left out of the analytic sample students who shared a certain characteristic measured prior to the introduction of the intervention, these exclusions do not compromise random assignment. Furthermore, the excluded subjects may be removed from the attrition calculation because they were based on a preintervention characteristic. This outcome is represented by the green box in figure II.3, and the distinction between this outcome and exclusions that are counted as attrition is discussed further in step 2 under the subsection on sample loss that is not considered attrition.

The WWC considers an RCT to be compromised only when the researcher analyzes data subject to one of these four concerns. Some valid randomization procedures can produce intervention and comparison groups that appear dissimilar based on chance. The WWC does not consider these chance differences to compromise the RCT, and such studies are reviewed using the usual review process for valid RCTs. Also, if a study reports multiple findings, only some of which the WWC determines to be compromised RCTs, then the findings that maintain the integrity of the random assignment can be reviewed using the process for valid RCTs.

QEDs

A study is eligible to be reviewed as a QED if it compares outcomes for subjects in an intervention group with outcomes for subjects in a comparison group but does not rely on random assignment to determine membership in the two groups. Groups can be identified through a variety of processes and be eligible for WWC review as long as the groups are exclusive, meaning a subject can be analyzed as a member of only a single group. Assignment to the intervention may depend on both observed and unobserved characteristics. For example, a group of students may be eligible for an afterschool program, but only some students may choose to participate. The students who did not choose to participate are designated as the comparison group. In this case, the characteristics of intervention and comparison groups differ. The two groups may differ on characteristics researchers were able to measure, such as test scores, or on characteristics that researchers were not able to measure, such as motivation. Even with equivalence on measured characteristics, there may be differences in unmeasured characteristics that could introduce *bias* into an estimate of the effect of the intervention. Bias is a systematic difference between the true impact of the intervention and the estimated impact, which can lead to incorrect conclusions about the effect of the intervention. For this reason, QEDs cannot receive the highest WWC rating but can receive the rating *Meets WWC Group Design Standards With Reservations*.

WWC review process for step 1 of the review of individual-level assignment studies

- ▶ If individuals have been placed into each study condition through a valid random assignment process and the RCT has not been compromised, then continue to [step 2](#).
- ▶ If individuals have *not* been placed into each study condition through a valid random assignment process but the study is eligible for review as a QED, then continue to [step 3](#).

Step 2. Sample attrition: Is the combination of overall and differential attrition high?

Attrition occurs when an outcome variable is not available for all subjects initially assigned to the intervention and comparison groups. Even well-designed RCTs may experience rates and patterns of sample attrition that compromise the initial comparability of the intervention and comparison groups, and potentially lead to biased estimates of the intervention's effectiveness. Attrition leads to bias when it is related to the outcome of interest. For RCTs, the WWC is concerned about both *overall attrition*—the rate of attrition for the entire sample, measured as the percentage of the randomized sample that has been lost—and *differential attrition*—the percentage point difference in the rates of attrition for the intervention and comparison groups. Both types of attrition contribute to the potential bias of the estimated effect.

High and low attrition

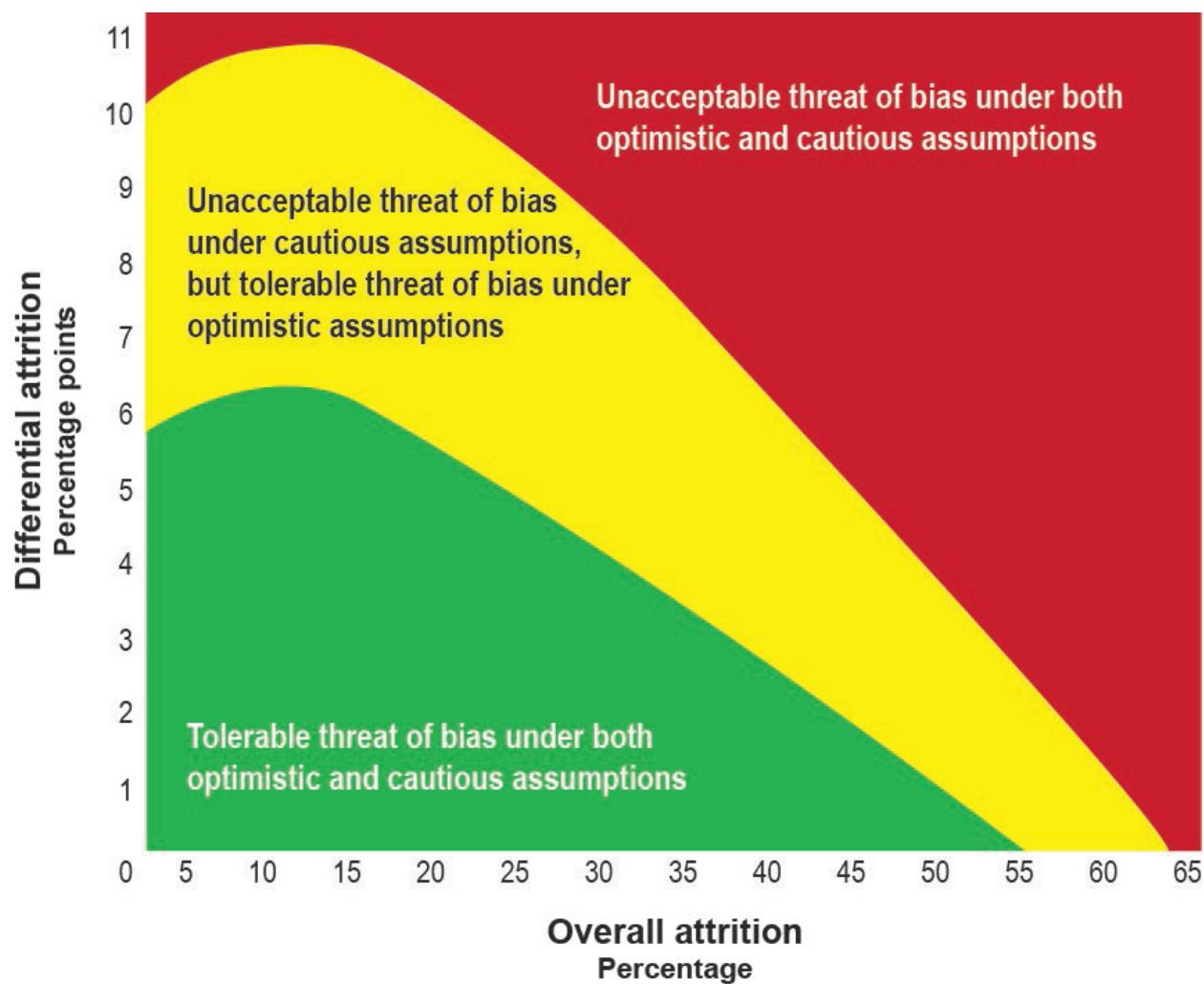
The WWC's attrition standard is based on a theoretical model for attrition bias and empirically based assumptions. The model depicts potential bias as a function of the rates of overall and differential attrition and the relationship between attrition and outcomes. To determine reasonable values to use in assessing the extent of potential attrition bias in a study, the WWC made assumptions about the relationship between attrition and outcomes that are consistent with findings from several randomized trials in education. More information on the model and the development of the attrition standard can be found in the WWC Technical Paper on [Assessing Attrition Bias](#).

Based on applying more *optimistic* or more *cautious* sets of assumptions about the relationship between attrition and outcomes to the model, the WWC has measured the levels of expected bias associated with different combinations of overall and differential attrition rates. Figure II.2 illustrates an approximation of the combinations that generate tolerable (green region), potentially tolerable (yellow region), and unacceptable (red region) levels of expected bias. A tolerable level of bias is defined as an effect size of 0.05 standard deviation or smaller on the outcome, which represents about 2 percentile points for a student scoring at the 50th percentile. For example, if the results reported in a study suggest the intervention will move the student from the 50th percentile to the 60th percentile, an effect size of 0.25 standard deviation, the actual impact of the intervention may move the student only to the 58th percentile, an effect size of 0.20 standard deviation. The WWC's threshold for the tolerable level of bias was based on extensive consultation with experts.

- The red region of figure II.2 shows combinations of overall and differential attrition that result in unacceptable levels of potential bias for both the optimistic and cautious sets of assumptions.
- The green region of figure II.2 shows combinations of overall and differential attrition that result in tolerable levels of potential bias for both the optimistic and cautious sets of assumptions.

Within the yellow region of the figure, whether the potential bias exceeds 0.05 standard deviation depends on the set of assumptions used. In developing the review protocol as described in the *Procedures Handbook*, the review team leadership considers the types of samples and the likely relationship between attrition and outcomes for studies in the area to guide their choice. Either the optimistic or cautious assumptions are chosen and specified in the review protocol to be applied consistently for all studies within the review.

- If the review team leadership has reason to believe that much of the attrition is exogenous to the interventions reviewed—that is, unrelated to the intervention—then more *optimistic* assumptions regarding the relationship between attrition and the outcome may be appropriate for a review. For example, the review team leadership may choose the optimistic assumptions if it believes attrition most likely arises from the movement of young children in and out of school districts due to family mobility or from typical absences on the days that assessments are conducted. In this case, the yellow region shows combinations that result in tolerable levels of potential bias, along with green.
- If the review team leadership has reason to believe that much of the attrition is endogenous—that is, related to the intervention—then more *cautious* assumptions may be appropriate for a review. For example, the review team leadership may choose the cautious assumptions for reviews of dropout prevention programs that rely on voluntary participation. In this case, the yellow region shows combinations that result in unacceptable levels of potential bias, along with red.

Figure II.2. Attrition and potential bias

Note: Not every combination of differential and overall attrition is possible for any given study. The review protocol will specify which set of attrition boundary values applies.

When the combination of overall and differential rates of attrition results in unacceptable levels of potential bias—the red region, along with the yellow region if making more cautious assumptions—the WWC labels this *high attrition*. When the combination of overall and differential rates of attrition result in tolerable levels of potential bias—the green region, along with the yellow region if making more optimistic assumptions—the WWC labels this *low attrition*. Therefore, the choice of optimistic or cautious assumptions results in a specific set of combinations of overall and differential rates of attrition that defines high and low attrition to be applied consistently for all studies in an area.

For each overall attrition rate, table II.1 shows the highest differential attrition rate allowable to still be considered low attrition under the two possible assumptions: cautious and optimistic.

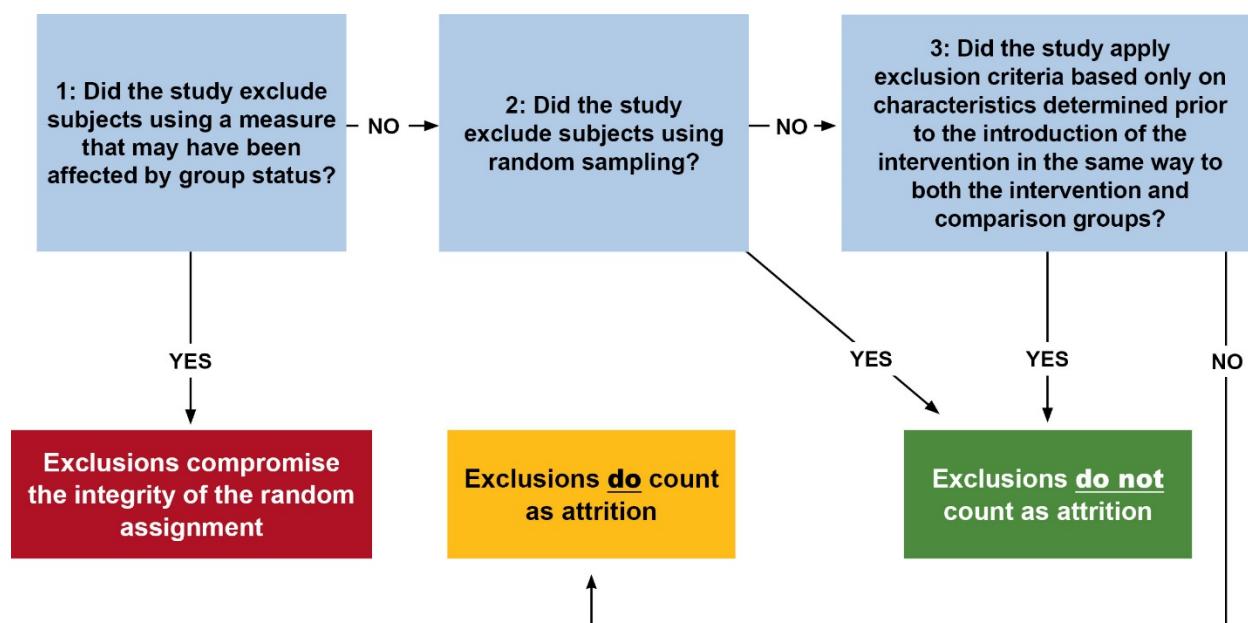
Sample loss that is not considered attrition

Not all loss of sample after random assignment is included in attrition calculations:

- Losing sample members after random assignment because of acts of nature, such as hurricanes or earthquakes, is not considered attrition when the loss is likely to affect intervention and comparison group members in the same manner. However, when sample loss due to an act of nature was concentrated in one group, the loss will be considered attrition.
- The excluded sample when analyzing outcome data for only a subset of the initial sample is not considered attrition if the subsample of the intervention or comparison group was randomly selected or if the subsampling was based on characteristics that were clearly determined prior to the introduction of the intervention and applied consistently across the intervention and comparison groups. For example, students who were excluded for having individualized education programs prior to the study would not be counted as attrition. The WWC considers characteristics that are unlikely to change over time, including sex and race or ethnicity, as having been determined prior to the introduction of the intervention, even when the researchers collected these data later.

The WWC presumes that sample loss arising from sources other than those described above could be related to outcomes and includes this sample loss in attrition calculations. The WWC's rules for how sample exclusions can affect the rating of an RCT are summarized in figure II.3. This includes sample exclusions that can compromise the RCT described under step 1 (red box), sample exclusions that are not considered attrition based on the criteria above (green box), and all other sample exclusions that are counted as attrition (yellow box).

Figure II.3. How the What Works Clearinghouse treats sample exclusions in randomized controlled trials



A characteristic can be determined after random assignment but not affected by group status, so the answers to the questions in boxes 1 and 3 can both be “no.” For example, the answer to all

three questions would be “no” in a study where the researcher excluded students with individualized education program statuses at a point in time after the introduction of the intervention, but the intervention is unrelated to how students are identified for individualized education programs. In this example, the excluded sample would be counted as attrition because some students’ statuses may have been determined after the introduction of the intervention but were not likely to be affected by the intervention. However, if there is a clear channel through which the intervention could affect the individualized education program status of students, then the sample exclusion would compromise the RCT. For example, the exclusion would compromise the RCT if the intervention provided support to teachers on identifying students who may need individualized education programs.

Table II.1. Highest differential attrition rate for a sample to maintain low attrition, by overall attrition rate, under “optimistic” and “cautious” assumptions

Overall attrition	Differential attrition		Overall attrition	Differential attrition		Overall attrition	Differential attrition	
	Cautious boundary	Optimistic boundary		Cautious boundary	Optimistic boundary		Cautious boundary	Optimistic boundary
0	5.7	10.0	22	5.2	9.7	44	2.0	5.1
1	5.8	10.1	23	5.1	9.5	45	1.8	4.9
2	5.9	10.2	24	4.9	9.4	46	1.6	4.6
3	5.9	10.3	25	4.8	9.2	47	1.5	4.4
4	6.0	10.4	26	4.7	9.0	48	1.3	4.2
5	6.1	10.5	27	4.5	8.8	49	1.2	3.9
6	6.2	10.7	28	4.4	8.6	50	1.0	3.7
7	6.3	10.8	29	4.3	8.4	51	0.9	3.5
8	6.3	10.9	30	4.1	8.2	52	0.7	3.2
9	6.3	10.9	31	4.0	8.0	53	0.6	3.0
10	6.3	10.9	32	3.8	7.8	54	0.4	2.8
11	6.2	10.9	33	3.6	7.6	55	0.3	2.6
12	6.2	10.9	34	3.5	7.4	56	0.2	2.3
13	6.1	10.8	35	3.3	7.2	57	0.0	2.1
14	6.0	10.8	36	3.2	7.0	58	-	1.9
15	5.9	10.7	37	3.1	6.7	59	-	1.6
16	5.9	10.6	38	2.9	6.5	60	-	1.4
17	5.8	10.5	39	2.8	6.3	61	-	1.1
18	5.7	10.3	40	2.6	6.0	62	-	0.9
19	5.5	10.2	41	2.5	5.8	63	-	0.7
20	5.4	10.0	42	2.3	5.6	64	-	0.5
21	5.3	9.9	43	2.1	5.3	65	-	0.3

Note: Overall attrition rates are given as percentages. Differential attrition rates are given as percentage points. Not every combination of differential and overall attrition is possible for any given study. The review protocol will specify which set of attrition boundary values applies.

Source: WWC Technical Paper on [Assessing Attrition Bias](#).

Measuring attrition when there is imputed outcome data

When a study is missing outcome data, researchers may replace the unobserved data with data that have been imputed in some way, rather than exclude subjects with missing outcome data from the analytic sample. Sample members with missing and then imputed data are considered to be missing when computing attrition. Using this approach, the result of the attrition calculation is the same regardless of how authors address the missing data. For example, if a study analyzes data from 100 subjects, including 90 with measured outcome data and the remaining 10 with outcome data imputed by the researchers, then the overall attrition rate is 10 percent. See section II.C for more information on how the WWC reviews studies with missing or imputed baseline or outcome data.

WWC review process for step 2 of the review of individual-level assignment studies

- ▶ If the RCT has a combination of overall and differential rates of sample attrition that meets the criteria for low attrition, then the study is eligible to *Meet WWC Group Design Standards Without Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the RCT has unknown or high levels of sample attrition, then continue to [step 3](#).

Step 3. Baseline equivalence: Is equivalence established at baseline for the groups in the analytic sample?

RCTs with high attrition, compromised RCTs, and all QEDs are ineligible to receive the highest WWC rating because of uncertainty about intervention and comparison group similarity prior to the introduction of the intervention. For these studies, equivalence of the intervention and comparison groups on specified characteristics measured at *baseline*—that is, prior to the introduction of the intervention—must be assessed for the *analytic sample*, the subjects from the intervention and comparison groups used to estimate findings. The characteristics on which the WWC must assess baseline equivalence are specified in the review protocol.

If the reported difference of a specified baseline characteristic is greater than 0.25 standard deviation in absolute value, based on the variation of that characteristic in the pooled sample of intervention and comparison group members, the WWC considers the intervention and comparison groups to be nonequivalent. For differences in the specified baseline characteristics that are between 0.05 and 0.25 standard deviation, the analysis must include an acceptable *statistical adjustment* for the baseline characteristics to meet the baseline equivalence requirement. Differences of less than or equal to 0.05 standard deviation require no statistical adjustment (table II.2). Chapter VI of the *WWC Procedures Handbook, Version 4.1*, describes the formulas the WWC uses to calculate these standard deviation differences, or effect sizes, for both continuous and dichotomous measures.

Table II.2. Absolute effect size at baseline

$0.00 \leq \text{Baseline ES} \leq 0.05$	$0.05 < \text{Baseline ES} \leq 0.25$	$ \text{Baseline ES} > 0.25$
Satisfies the baseline equivalence requirement	Requires statistical adjustment to satisfy the baseline equivalence requirement	Does not satisfy the baseline equivalence requirement

ES is effect size.

The statistical adjustments the WWC considers acceptable depend on the relationship between the baseline characteristic and the outcome. In general, when the WWC requires an analysis to include a statistical adjustment for a baseline characteristic specified in the review protocol, the characteristic must be included in the analysis at the subject level such that it accounts for the correlation between the baseline measure and the outcome. Several techniques are acceptable to meet this requirement, including regression adjustment and analysis of covariance.

However, when the baseline characteristic is the same as the outcome, additional approaches that do not estimate a correlation may also be acceptable. These methods include using simple gain scores, applying a difference-in-differences adjustment, or including individual-level fixed effects.¹ If the authors do not perform the adjustment themselves, then the WWC can perform its own difference-in-differences adjustment, as described in appendix E of the *WWC Procedures Handbook, Version 4.1*, to allow the study to satisfy the statistical adjustment requirement. The WWC will consider these additional approaches as acceptable statistical adjustments if the following two conditions are met:

1. **The baseline and outcome measures must be measured using the same units.** For example, this condition would be satisfied if the researchers administered the same test, using the same scoring procedures, as a pretest and posttest. This condition would not be satisfied if (a) the researchers administered different assessments at baseline and follow-up or (b) the measures were the same, but different subscales or scoring procedures were used to score the tests.
2. **The baseline characteristic must have a correlation of .60 or higher with the outcome.** In general, the correlation must be estimated using the study data. However, topic areas may waive this requirement for a measure or outcome domain if the protocol documents evidence that the correlations between pretests and posttests of the measure typically exceed .60, and the exception is applied consistently for all studies within the review.

The review protocol can also specify a maximum elapsed time between the assessment of the baseline and outcome measures used in these approaches. Importantly, these requirements must only be met when the approach is used to satisfy the WWC's statistical adjustment requirement in a study with a baseline difference between 0.05 and 0.25 standard deviation.

The approaches the WWC considers acceptable are summarized in table II.3. Additional considerations for statistical adjustments in some common analytic approaches, including propensity score analyses and analyses in which subjects are observed in multiple time periods, are described in section II.C.

When the WWC does not require a statistical adjustment (because the study is a low-attrition RCT or has baseline differences less than or equal to 0.05 standard deviation), authors can adjust their analyses using approaches besides those that the WWC considers acceptable for the purpose of satisfying the statistical adjustment requirement. Furthermore, although the WWC standards require statistical adjustments in limited circumstances and only for certain specified characteristics, authors may adjust for all available baseline data in their analyses.

¹ A difference-in-differences adjustment involves subtracting the baseline difference from the difference in outcomes measured at follow-up.

Table II.3. Examples of acceptable approaches for satisfying the What Works Clearinghouse statistical adjustment requirement

Acceptable methods for any baseline measure
<ul style="list-style-type: none"> • Regression covariate adjustments in ordinary least squares models. • Regression covariate adjustments in hierarchical linear models. • Analysis of covariance. • Other approaches to regression covariate adjustments, including nonlinear regression analysis, such as logistic or probit models.
Acceptable methods when the baseline and outcome measures are the same and have a strong relationship ^a
<ul style="list-style-type: none"> • Simple gain scores. • Difference-in-differences adjustment. • Fixed effects for individuals.

^aTwo conditions must hold for these to be considered acceptable statistical adjustments for a baseline and outcome measure that are the same: the baseline and outcome measures must be measured using the same units and the baseline characteristic must have a correlation of .60 or higher with the outcome.

Additional considerations regarding assessing and satisfying the baseline equivalence requirement are as follows:

- Baseline equivalence must be assessed separately for each analytic sample. Satisfying the baseline equivalence requirement on one analytic sample does not positively or negatively affect the requirement for other analytic samples, even for outcome measures in the same domain. For example, consider a QED that measured impacts using both the full sample and a sample that excluded one student. In this example, it is necessary to assess baseline equivalence on each sample separately.
- Preintervention measures used to assess baseline equivalence must satisfy the same reliability criteria specified for outcomes, as described in chapter V. If reliability information for a preintervention measure is required, but unavailable, or if the reliability is below the acceptable level, then the measure cannot be used to assess baseline equivalence.
- A baseline measure assessed after the start of the intervention can be used to satisfy the baseline equivalence requirement. However, if a significant portion of the intervention occurred prior to the assessment of a baseline measure used to satisfy the baseline equivalence requirement, then the WWC will note in its reporting that the study measures the effect of the portion of the intervention that occurred after the measure was assessed and until the time of the follow-up assessment. If both preintervention and intermediate measures are available, then the WWC will use the preintervention measure to assess baseline equivalence.
- When the WWC requires a statistical adjustment to satisfy the baseline equivalence requirement, the study must report the direction, but not necessarily the magnitude, of the impact estimate from the analysis that includes the required statistical adjustment. For example, the authors of a study might perform an acceptable adjustment, but the WWC may be unable to obtain the information needed to measure the magnitude of the finding or calculate its effect size. For this study to be eligible to be rated *Meets WWC Group Design Standards With Reservations*, it must report the direction of the adjusted finding. If the

authors do not provide any information about the direction of the adjusted finding, then the study is rated *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards.

- Section II.C discusses additional considerations for assessing baseline equivalence in studies with missing or imputed data. First, while all QEDs must satisfy the baseline equivalence requirement, high-attrition RCTs that impute outcome data and analyze the full sample that was randomized to conditions do not need to satisfy the baseline equivalence requirement to be eligible to be rated *Meets WWC Group Design Standards With Reservations*, as described in step 4 of section II.C. Second, if an analytic sample includes missing or imputed data for a specified preintervention measure, then it must satisfy the baseline equivalence requirement using the largest baseline difference under different assumptions about how the missing data are related to measured or unmeasured factors, as described in step 5 of section II.C. Finally, all studies must use one of the acceptable approaches listed in table II.6 in section II.C to address missing data in the analytic sample to be eligible to be rated *Meets WWC Group Design Standards With or Without Reservations*.
- If the study used weights in the analysis, then the baseline means must also be calculated using the same weights.
- If the study conducted random assignment within blocks or matching within strata, and the analysis includes dummy variables that differentiate these blocks or strata, then the baseline means may also be adjusted using these same dummy variables (Wolf, Price, Miller, & Boulay, 2017).

Some additional considerations provide review teams with discretion in how the baseline equivalence requirement is satisfied. Discretion is needed because the outcome measures and outcome *domains*—sets of closely related outcomes—used in different reviews can vary substantially. When the review team leadership exercises discretion, the approach must be specified in the review protocol and applied consistently for all studies within the review. These additional considerations include the following:

- Baseline equivalence must be assessed separately for each outcome domain. The review protocol will describe eligible outcome domains and specify which preintervention measures can or must be used to satisfy the baseline equivalence requirement for each. Unless specified in the protocol, satisfying the baseline equivalence requirement in one domain does not affect the requirement in other domains.
- When the outcome measure is a test of academic achievement, the review protocol often specifies that baseline equivalence must be assessed using a preintervention measure of academic achievement. However, for outcome measures that cannot be measured at baseline, such as completing high school, the review protocol will instead specify background characteristics—such as academic achievement, socio-economic status, or other measures that are related to the outcome of interest—on which baseline equivalence must be assessed.
- When specifying the preintervention measures used to satisfy the baseline equivalence requirement within a domain, the protocol can list those in the same or different domain from the outcome measure. For example, free or reduced-price lunch status might be required for satisfying baseline equivalence in a domain for measures of staying in

school. Additionally, preintervention measures that are related to but are different from the outcome are typically allowed. For example, a study might examine impacts on a state-administered standardized test at the end of grade 3 but report on a researcher-developed measure that covers similar content at the beginning of grade 3. Depending on the protocol, the researcher-developed measure could be used to satisfy the baseline equivalence requirement.

- When preintervention measures in a particular domain are thought to have strong relationships with outcome measures in all domains within a topic area, review protocols may specify domains on which baseline equivalence must be assessed even when the study does not report on findings for outcomes in the domain. For example, a review protocol for a behavior-focused topic area might require baseline equivalence on a preintervention measure of behavior even for academic achievement outcomes.
- A difference larger than 0.25 standard deviation for *any* specified preintervention measure in a domain means that *all* of the outcomes in the domain fail to satisfy the baseline equivalence requirement because domains are typically defined to include outcomes that are thought to be highly correlated. However, review protocols with eligible domains that include a broader set of outcomes may require instead that equivalence be assessed outcome-by-outcome rather than domain-by-domain. In the outcome-by-outcome approach to baseline equivalence, baseline equivalence for each outcome measure is assessed using a pretest of the outcome, and that pretest does not positively or negatively affect the requirement for other outcome measures in the same domain. However, the review protocol using the outcome-by-outcome approach for a domain must specify whether it is possible to satisfy the baseline equivalence requirement for an outcome measure when a pretest is not available, but a different related measure was assessed at baseline.
- When a study reports findings for multiple outcome measures within a domain, the WWC requires that analyses of all outcomes in that domain include statistical adjustments for all preintervention measures that require adjustment in that domain. For example, if A, B, and C are available as preintervention measures and outcomes for the same analytic sample, and the preintervention difference for B requires statistical adjustment, then the preintervention measure of B must be included for each of the analyses of A, B, and C. In the case of a review protocol with eligible domains that include a broad set of outcomes, the protocol may instead require that the adjustment for a preintervention measure of an outcome measure be included only for that outcome measure.

WWC review process for step 3 of the review of individual-level assignment studies

- If the study satisfies the baseline equivalence requirement for the analytic intervention and comparison groups on the characteristics specified in the review protocol—including acceptable statistical adjustments, if necessary—then the study is eligible to receive the rating *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the study does not satisfy the baseline equivalence requirement for the analytic intervention and comparison groups on the characteristics specified in the review protocol, then the study is rated *Does Not Meet WWC Group Design Standards*.

B. Cluster-level assignment

Research studies in which individuals are assigned to the intervention or comparison condition as groups, known as *clusters*, have become more common in education research. This cluster-level assignment can take a number of forms, including students grouped within teachers, students grouped within classrooms, students grouped within schools, teachers grouped within schools, or classrooms grouped within schools.

Studies may involve random assignment of clusters but use individual-level information within those clusters to estimate impacts. In these studies, the observed effects of the intervention can be influenced both by the effects of the intervention on individuals and by changes in the composition of individuals within clusters. For example, a highly attractive intervention may draw students from other classrooms or schools between the time of random assignment and when outcomes are measured. The WWC reviews cluster-level assignment studies to determine whether the observed effects of the intervention can be credibly said to be due solely to the intervention's *effects on individuals*, or whether changes in the composition of individuals may also have affected the findings. If compositional changes cannot be ruled out, then the study may still satisfy WWC standards for evidence of the intervention's *effects on clusters* but cannot achieve the highest WWC rating.²

Some cluster-level assignment studies analyze individual-level outcomes and others analyze cluster-level outcomes—that is, individual-level outcomes that have been aggregated to the cluster level, but the distinction between an intervention's effects on clusters and its effects on individuals is not based on the unit of analysis. It is possible for an analysis of cluster-level data to satisfy WWC standards for evidence of effects on individuals, and similarly it is possible for an analysis of individual-level data to satisfy WWC standards for evidence of effects on clusters.

This section presents criteria under which estimates of effects from cluster-level assignment studies can be rated *Meets WWC Group Design Standards Without Reservations*, *Meets WWC Group Design Standards With Reservations*, or *Does Not Meet WWC Group Design Standards*. Figure II.4 displays the steps for determining a cluster-level assignment's rating. The WWC initially reviews the evidence of an intervention's effect on individuals (steps 1–4). If an effect on individuals cannot be credibly demonstrated, then the WWC reviews the evidence of an intervention's effect on clusters (steps 5–7), where changes in the composition of individuals within the clusters may influence the observed effect. Each step involves addressing a question about the study's research design. The answer to each question leads to subsequent steps that should be taken as part of the review process (figure II.4). In the steps that follow, assessments of attrition and baseline equivalence will use the same standards described in section II.A, with some noted exceptions.

Cluster-level assignment studies that satisfy the requirements outlined in steps 1 to 7 are eligible to be rated *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*. However, to receive one of these ratings, the study must also satisfy the requirements in chapter V, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

² Although measuring the effects of an intervention on schools or other clusters can answer important policy questions, the WWC focuses primarily on evidence of effects on students, or on outcomes for educators thought to be relevant to improving outcomes for students.

Screening criteria to determine whether the study is a cluster-level assignment study

A study should be reviewed using the standards for cluster-level assignment studies when it satisfies two conditions: individuals are assigned to the intervention or comparison condition as groups and outcomes are measured for individuals within those clusters and may be analyzed as individual-level data or as cluster-level averages.

Based on these two criteria, neither the method of impact estimation nor the level of aggregation of data determines whether the study should be considered an individual-level or cluster-level assignment study. Consider a study that randomly assigns schools to a condition; the outcome of interest is student achievement, but the data are aggregated to the school level for the analysis of average achievement levels by school. The study meets the first condition because it assigned schools to conditions. The study meets the second condition because the outcome measure was assessed for individuals within schools, and it does not matter that the study aggregated the data to the school level. Put another way, this study would still be considered a cluster-level assignment study, even though the unit of analysis is aligned with the unit of assignment, namely the school, because the aggregated data actually represent outcomes measured at the individual level.

We provide three additional examples of the application of these screening criteria in RCTs:

- If a study randomly assigns teachers to a condition, and the outcome of interest is a student outcome, such as achievement, then the study should be characterized as a cluster-level assignment study. The unit of assignment is the teacher, and the outcome was measured for individual students, whether aggregated to the teacher level for analysis or not.
- If a study randomly assigns teachers to a condition and the outcome of interest is a teacher outcome, such as retention, then the study should be characterized as an individual-level assignment study. The unit of assignment is the teacher, and the outcome was measured for teachers.
- If a study randomizes both clusters and individuals, then the study is an individual-level assignment study. For example, a study might randomize classrooms to a condition and randomize students to classrooms, in either order. In this case, the unit of assignment is the student.

The two screening criteria—individuals are assigned to condition as groups and outcomes are measured for individuals within clusters—also apply to QEDs, and the study description may provide guidance regarding the appropriate unit of assignment. For example, “schools using the intervention were compared against schools not using the intervention” illustrates a situation in which the cluster, in this case the school, is the unit of assignment. Review protocols may also clarify how to identify the unit of assignment in scenarios common to a topic area. Otherwise, the WWC identifies the largest study unit that contains only members of one condition. For example, if a study examines the effect of an intervention on student achievement within a school and each classroom has *only* intervention students or *only* comparison students, then the unit of assignment is the cluster. In contrast, if some classrooms have both intervention and comparison students, then the unit of assignment is the individual. Similarly, a study that examined the effect of a dropout prevention program by comparing school-level dropout rates in intervention schools with the rates in comparison schools is a cluster-level assignment study. The unit of assignment is the school, and the outcome was measured for students.

Findings in a study that meet these two screening criteria could be influenced by changes in the composition of individuals within clusters and should be considered a cluster-level assignment study and reviewed using the following steps. If a group design study does not meet both criteria, then it should be reviewed as an individual-level assignment study.

Process for reviewing evidence of an intervention's effect on individuals (steps 1–4)

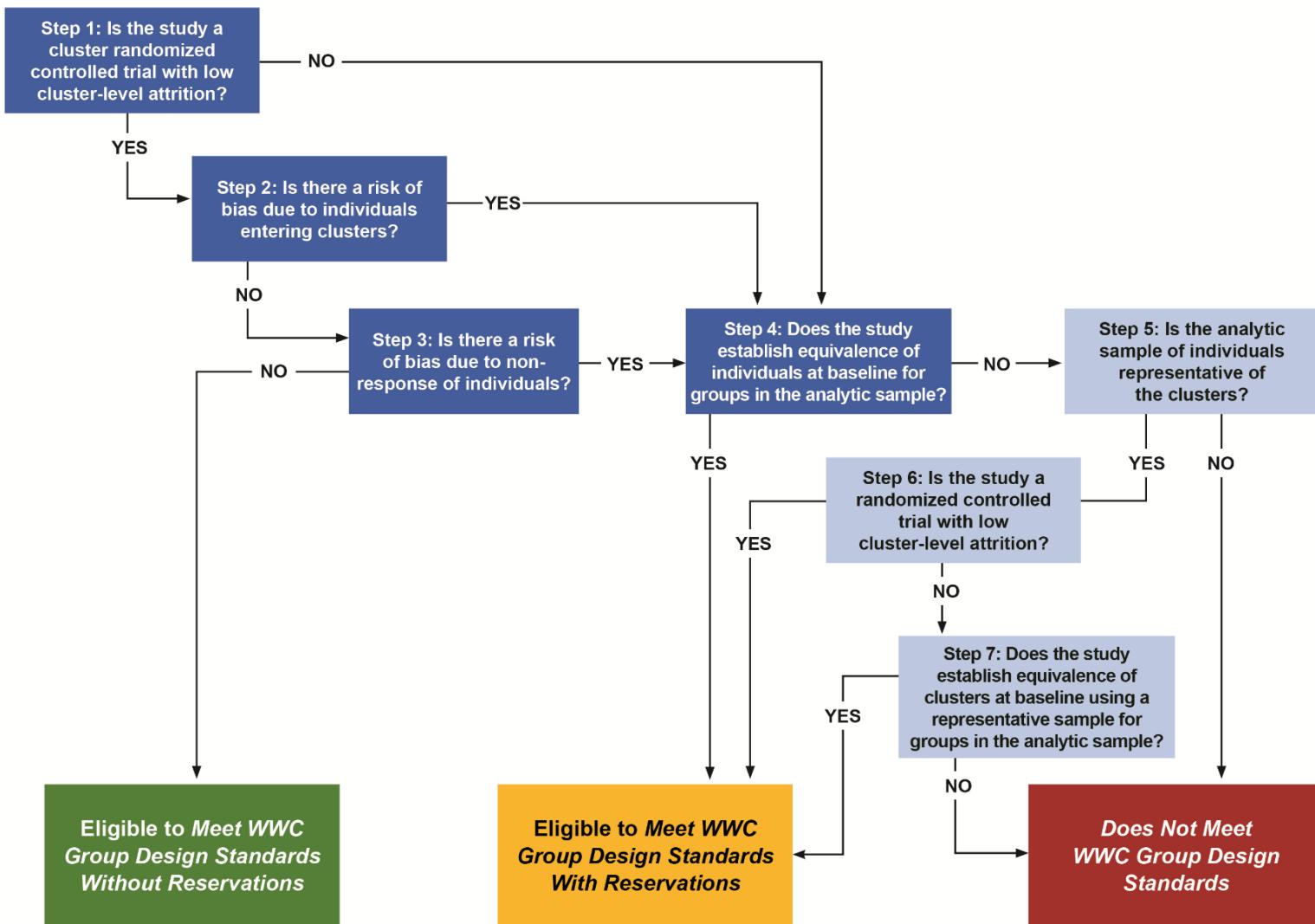
The following four steps describe the review process to assess the credibility of the evidence in a study for understanding the effects of an intervention on individuals. To be eligible to be rated *Meets WWC Standards Without Reservations*, a cluster RCT must limit potential bias from changes in the composition of clusters and individuals within clusters after random assignment. Cluster RCTs that have a risk of bias from these compositional changes, and all cluster QEDs, can still be eligible to be rated *Meets WWC Standards With Reservations* in the review to satisfy WWC standards for evidence of effects on individuals if the study satisfies a requirement for the baseline equivalence of individuals in the analytic intervention and comparison groups. A study can satisfy WWC standards for effects on individuals regardless of the level of aggregation of data used in the analysis. In particular, findings based on an analysis of student achievement data aggregated to the school level can satisfy WWC standards for evidence of the effects of an intervention on students if it satisfies the requirements in steps 1 to 4.

Step 1. Is the study a cluster RCT with low cluster-level attrition?

In order to receive the highest rating, the study must be an RCT that assigned clusters to a condition and has low cluster-level attrition, as defined by the boundaries specified in the applicable review protocol and displayed in figure II.2 and table II.1. Cluster-level attrition measures the loss of entire clusters from the randomized sample. A cluster is lost when it contributes no outcome data to the analytic sample. The loss of individuals from within clusters is assessed in step 3.

WWC review process for step 1 of the review of cluster-level assignment studies

- ▶ If the study is an RCT with low cluster-level attrition, then continue to [step 2](#).
- ▶ If the study is an RCT with high or unknown cluster-level attrition, a compromised RCT, or a QED, then continue to [step 4](#).

Figure II.4. Review process for cluster-level assignment studies

Note:

To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in chapter V, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Step 2. Is there a risk of bias due to individuals entering clusters?

In order to receive the highest rating, a cluster RCT must limit the risk of bias due to *joiners*, that is, individuals who enter the cluster after the time of random assignment. If the study includes joiners in the analytic sample, then the estimate of the effect of the intervention on individual outcomes could be biased if the individuals who entered intervention clusters differ systematically from those who entered comparison clusters. This risk of bias may vary across substantive areas and interventions, and based on how long after random assignment the joining occurred. Therefore, the review protocol will identify groups of joiners who would pose a risk of bias if included in the analytic sample for a cluster RCT based on when they joined clusters, features of the intervention, and the unit of assignment. The approach must be specified in the review protocol and applied consistently for all studies within the review.

Some joiners may enter clusters after random assignment but before the individuals knew the randomly assigned conditions of the clusters. The WWC never considers these joiners to pose a risk of bias because the decisions that led these individuals to join clusters could not have been affected by the intervention. The burden for demonstrating that individuals could not have known about the intervention rests with the study authors. For example, random assignment of schools might occur over the summer prior to the start of a school year, but the intervention was not announced to families in the district until after the school year began. Students who joined schools prior to the announcement would not pose a risk of bias. Students who joined schools after the announcement might pose a risk of bias, depending on which of the following three options is specified in the review protocol:

- *When all joiners who enter clusters after the results of random assignment are known pose a risk of bias.* Some reviews may include studies of programs or policies that are likely to affect enrollment or placement decisions, such as school turnaround interventions that close or combine schools, or a policy that allows students to leave neighborhood schools for choices throughout the district. In these types of studies, joiners who enter intervention schools at any time after the results of random assignment are known may be different from joiners who enter comparison schools because they may choose the school for a specific reason. For example, if high-performing students view the intervention schools as better suited for them and switch into those schools after the study begins, then the observed effect may be biased by differences in the types of joiners who entered the schools. In this case, including any joiners in the analytic sample who enter schools after the results of random assignment are known would pose a risk of bias because students or their families may choose the intervention schools for reasons specifically related to the intervention. Other cases when all joiners may pose a risk of bias include when classrooms or teachers are assigned to conditions, but students are non-randomly assigned to classrooms later by the principal or other school personnel. If those responsible for assigning students to classrooms exercise discretion for reasons specifically related to the intervention, then the observed effect may be biased by differences in students assigned to the intervention and comparison groups.
- *When only late joiners pose a risk of bias.* Some reviews may include studies in which *early joiners*, students who enter a school soon after the study begins are not likely to be a source of bias, but *late joiners*, students entering later, may be. For example, schools may be randomly assigned to implement a reading supplement or professional development program prior to or at the beginning of a school year. Some students may enter a school as

the school year begins or shortly after because they have just moved to the neighborhood, which is common in many school settings. These early joiners are unlikely to have chosen the school for reasons related to the intervention because the intervention is just beginning in the school and little may be known about it. Therefore, those early joiners may not differ from students who enter comparison schools early in the school year. However, students who enter schools later in the school year may be more likely to do so because of the intervention, and therefore these students might differ from those who enter comparison schools later. As a default, early joining is defined as occurring within the first six weeks of the school year, but a different length for this initial period that differentiates between early or late joiners can be specified in the review protocol.³

- *When no joiners pose a risk of bias.* Some reviews might focus on settings in which there is little or no risk of bias from individuals who enter clusters at any point after initial random assignment. For example, interventions that have a very low profile, such as a change to recess programming or a low-profile teacher mentoring program, would not be expected to represent a significant draw for students, so individuals who join intervention clusters are likely to be similar to those who join comparison clusters. In these instances, the review protocol may specify that individuals who enter clusters after the results of random assignment are known may be included in the analytic sample without a risk of bias.

The review protocol may select different options for which joiners pose a risk of bias for different groups of interventions and for different units of assignment. For example, the review protocol might indicate that no joiners pose a risk of bias when the unit of assignment is the school, but all joiners pose a risk of bias when the unit of assignment is the classroom, teacher, or smaller unit.

Table II.4 summarizes the categories of interventions that fall into each of these three categories.

Table II.4. Three categories of joiner risk specified in review protocols

All joiners after the results of random assignment are known pose a risk of bias.	<ul style="list-style-type: none"> • Appropriate for interventions that are likely to influence placement or enrollment. • Joiners who enter intervention clusters are likely different from those who enter comparison clusters.
Late joiners pose a risk of bias.	<ul style="list-style-type: none"> • Appropriate for studies in which joiners who enter soon after the study begins are not likely to be a threat, but later joiners may be. • Early joiners are as good as randomly assigned, but late joiners are a concern.
No joiners pose a risk of bias.	<ul style="list-style-type: none"> • Appropriate for interventions with a very low profile. • Both early and late joiners are as good as randomly assigned.

³ For interventions that begin later in the school year, the review team leadership may judge whether the joiners pose a risk of bias similar to those of early or late joiners as defined in the review protocol.

A study that excludes all joiners from the analytic sample or only includes joiners who do not pose a risk of bias is said to *limit the risk of bias from joiners* and is eligible to be rated *Meets WWC Group Design Standards Without Reservations* if the study also has low levels of individual-level nonresponse (step 3). However, if a study includes any joiners in the analytic sample who pose a risk of bias according to the review protocol, then the highest rating the study can receive is *Meets WWC Group Design Standards With Reservations*. To receive that rating in the review for evidence of effects on individuals, the study must satisfy the baseline equivalence requirement on the characteristics specified in the review protocol for the individuals in the analytic intervention and comparison groups (step 4).

WWC review process for step 2 of the review of cluster-level assignment studies

- ▶ If the study either excludes all joiners from the analytic sample or includes joiners in the analytic sample who do not pose a risk of bias, in accordance with the review protocol, then the study limits the risk of bias from joiners. Continue to [step 3](#).
- ▶ If the study's analytic sample includes joiners who entered after the results of random assignment are known and, in accordance with the review protocol, pose a risk of bias, then continue to [step 4](#).

Step 3. Is there a risk of bias due to nonresponse of individuals?

To receive the highest rating, a cluster RCT with a limited risk of bias due to joiners must have low individual-level nonresponse as well as low cluster-level attrition, assessed in step 1. Nonresponse at the individual level for cluster-level assignment studies is the difference between the individuals present in a reference sample, as described in table II.5, and those present in the analytic sample at the time the outcome is assessed. For studies that analyze outcomes aggregated to the cluster level, the individuals present in the analytic sample consist of those who contribute data to the outcome measure. The reference sample—the benchmark sample from which nonresponse is measured—can differ depending on the risk of bias associated with joiners. When the reference sample is the original randomized sample, this step measures individual-level *attrition*. Because the reference sample can differ from the randomized sample, the WWC refers to this step as measuring individual-level *nonresponse*.

Individual-level nonresponse is always measured within the sample of nonattriting clusters. Individuals in clusters not represented in the analytic sample do not contribute to the reference sample used in the denominator of the individual-level nonresponse calculation.

Table II.5. Allowable reference samples for calculating individual nonresponse

Joiners associated with a risk of bias as specified in protocol	Allowable reference samples
All joiners after the results of random assignment are known pose a risk of bias	1. Individuals present in nonattriting clusters prior to the announcement of the intervention
Only late joiners pose a risk of bias	Either (1), or 2. Individuals present in nonattriting clusters in early period
No joiners pose a risk of bias	Either (1), (2), or 3. Individuals in nonattriting clusters at follow-up

An acceptable reference sample must be defined at a point in time after all joiners included in the analytic sample had already joined clusters but before the time period associated with joiners

that pose a risk of bias according to the review protocol (table II.5). The first part of this requirement ensures that the reference sample includes all individuals contributing to the analytic sample. The second part of this requirement is based on the type of joiners that the review protocol specifies as posing a risk of bias, as follows:

- When *all joiners pose a risk of bias* except those who enter clusters before the results of random assignment are known, the only acceptable reference sample is (1) the sample of individuals who were present in nonattriting clusters at a point in time prior to the announcement of the intervention, for example, students in clusters at or before the time of random assignment.
- When *only late joiners pose a risk of bias*, the reference sample can be (1) or (2) the sample of individuals present in nonattriting clusters at a point in time within an initial early joiner period defined in the review protocol, for example, the rosters of students obtained from a school early in the first school year when the intervention was implemented.
- When no joiners pose a risk of bias, the reference sample can be (1), (2), or (3) the sample of individuals in nonattriting clusters at follow-up, for example, the number of students enrolled in study schools on the day the posttest was given.

If a study provides information for multiple acceptable reference samples for assessing individual-level nonresponse, the WWC will base its calculations on the earliest sample. Like the assessment of cluster-level attrition, the assessment of individual-level nonresponse will follow the boundaries specified in the applicable review protocol and displayed in figure II.2 and table II.1.

WWC review process for step 3 of the review of cluster-level assignment studies

- If the study has low levels of individual-level nonresponse, then it is eligible to be rated *Meets WWC Group Design Standards Without Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the study has high levels of individual-level nonresponse, then continue to [step 4](#).

Step 4. Does the study establish equivalence of individuals at baseline for groups in the analytic sample?

Cluster RCTs with a high risk of bias from attrition of clusters, inclusion of joiners in the analytic sample, or individual-level nonresponse, and cluster QEDs that satisfy the baseline equivalence requirement for the analytic sample of individuals, are eligible to be rated *Meets WWC Group Design Standards With Reservations*. The individuals in the analytic intervention and comparison groups must satisfy the same requirements specified in the baseline equivalence step of the review of individual-level assignment studies described in step 3 of section II.A. For studies that analyze outcomes aggregated to the cluster level, the individuals contributing data to the outcome measure must satisfy this requirement. Regardless of the level of analysis, this baseline equivalence requirement must be satisfied using individual-level standard deviations. Means calculated using either cluster- or individual-level data are acceptable as long as the weighting is consistent with the weighting used in the analysis. In general, a required statistical adjustment must be made using individual-level data, such that it accounts for the individual-level correlation between the baseline measure and the outcome. However, as in individual-level

assignment studies, using simple gain scores, applying a difference-in-differences adjustment, or including individual-level fixed effects are also acceptable approaches for required statistical adjustments when the baseline characteristic is measured using the same units as the outcome and the baseline characteristic has a correlation of .60 or higher with the outcome, using individual-level data to measure the correlation.

WWC review process for step 4 of the review of cluster-level assignment studies

- ▶ If the study satisfies the baseline equivalence requirement for the analytic sample of individuals on the characteristics specified in the review protocol, then it is eligible to be rated *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study does not satisfy the baseline equivalence requirement for the analytic sample of individuals, then it will be reviewed to determine whether it can satisfy WWC standards for the intervention's effect on clusters. This review process is described in the following section, beginning with step 5.

Process for reviewing evidence of an intervention's effect on clusters (steps 5–7)

The following three steps describe the review process to assess the credibility of the evidence in a study for understanding the effects of an intervention on clusters. In these studies, the observed impact estimate potentially represents a combination of the effect of the intervention on individuals and a composition effect due to different types of individuals entering intervention and comparison clusters. Therefore, evidence reviewed in this section is only eligible to be rated *Meets WWC Group Design Standards With Reservations*. To receive this rating, a study that did not receive a rating of *Meets WWC Group Design Standards With or Without Reservations* under the review to satisfy WWC standards for effects on individuals must analyze individuals who are representative of the clusters in the analytic sample, and the studies must be a cluster RCT with low cluster-level attrition or must satisfy a requirement for the baseline equivalence of clusters in the analytic intervention and comparison groups. A study can satisfy WWC standards for effects on clusters regardless of the level of aggregation of data used in the analysis. In particular, findings based on an analysis of individual-level student achievement data can satisfy WWC standards for evidence of the effects of an intervention on clusters if it satisfies the requirements in steps 5 to 7.

Step 5. Is the analytic sample of individuals representative of the clusters?

If a study has poor response rates at follow-up or sufficiently differential response rates among individuals in the intervention and comparison clusters, then the observed impact would not credibly estimate the effect of the intervention on clusters. Therefore, the WWC assesses the degree to which the individuals within clusters included in the analytic sample are *representative* of all individuals present in the clusters at follow-up.

The WWC assumes that findings based on administrative data satisfy the representativeness requirement unless review team leadership concludes that relevant individuals are excluded in a nonrandom way from those data. In all other cases, the WWC will assess representativeness using the attrition boundaries specified in the applicable review protocol and displayed in figure II.2 and table II.1. Like the calculation for individual nonresponse, only individuals in nonattriting clusters are counted. The numerator for this attrition calculation will be the number of individuals present in nonattriting clusters at follow-up—that is, at the approximate time when

outcomes were measured—who do not contribute to the analytic sample. The denominator for the attrition calculation will be the total number of individuals in nonattriting clusters at follow-up. Unlike the measurement of individual-level nonresponse in step 3, the reference sample in the denominator for measuring representativeness is always taken at follow-up. However, the timing of the reference sample count needs not be precisely aligned with the measurement of outcomes. For example, in a school-level assignment study that measured outcomes at the end of the school year, the reference sample might be the administrative school enrollment count taken at some point during the school year.

For studies that analyze outcomes aggregated to the cluster level, the individuals present in the analytic sample consist of those who contribute data to the outcome measure. The representativeness requirement is assessed using counts of individuals pooled across all clusters in the intervention or comparison group, not for each cluster individually.

WWC review process for step 5 of the review of cluster-level assignment studies

- ▶ If the study relies on administrative data or has low individual-level nonresponse for this representativeness assessment, then continue to [step 6](#).
- ▶ If the study has high or unknown individual-level nonresponse for this representativeness assessment, then it is rated *Does Not Meet WWC Group Design Standards*.

Step 6. Is the study an RCT with low cluster-level attrition?

This is the same assessment of cluster-level attrition from step 1. This step is repeated because it is possible for RCTs with low cluster-level attrition, RCTs with high cluster-level attrition, and QEDs to arrive at step 6.

WWC review process for step 6 of the review of cluster-level assignment studies

- ▶ If the study is an RCT with low levels of cluster attrition, then it is eligible to be rated *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study is an RCT with high or unknown cluster attrition or a QED, then move to [step 7](#).

Step 7. Does the study establish equivalence of clusters at baseline for groups in the analytic sample?

Among studies that did not receive a rating of *Meets WWC Group Design Standards With or Without Reservations* under the review to satisfy WWC standards for effects on individuals, those that are cluster RCTs with high or unknown cluster-level attrition and cluster QEDs must satisfy the baseline equivalence requirement for the analytic sample of intervention and comparison group clusters for the study to be eligible to be rated *Meets WWC Group Design Standards With Reservations*. The analytic sample of clusters consists of the clusters represented in the sample used to estimate findings.

The characteristics on which the WWC must assess baseline equivalence of clusters are specified in the review protocol and may differ from those used to assess baseline equivalence of individuals. Examples of characteristics include student achievement levels, grade levels, demographics of teachers or students in schools, and school setting. The review protocol will

also specify whether individuals contributing baseline data in the clusters used to assess baseline equivalence of clusters must be the same individuals contributing outcome data to the analysis.

In particular, the review protocol will determine the following parameters for satisfying the baseline equivalence requirement:

- **Whether the baseline equivalence requirement can be met using data from an earlier assessment of the same cohort of individuals in the analytic sample within the same clusters.** For example, for school-level assignment studies, a protocol may allow the requirement to be satisfied for an analytic sample of grade 4 students in 2015 using the same cohort in grade 3 in 2014. Although the same schools contribute outcome and baseline data, the students contributing baseline and outcome data will overlap but may not be identical because some students will transfer into or out of the schools between the two school years.
- **Whether the baseline equivalence requirement can be met using data from an earlier cohort of students within the same clusters.** For example, for school-level assignment studies, a protocol may allow the requirement to be satisfied for an analytic sample of grade 4 students in 2015 using grade 4 students in 2014 within the same schools. Aside from students who may have repeated grade 4, the students contributing baseline data are not the same as those contributing outcome data.
- **The maximum elapsed time that is allowed between the collection of baseline and outcome data when the individuals contributing baseline and outcome data are not identical.** As more time elapses between the collection of baseline and outcome data, the relevance of the baseline data may become weaker. For example, if outcomes are measured for grade 5 students in 2015 but baseline data are collected for the same cohort in grade 1 in 2011, there may be less overlap in the samples than if the baseline data were collected in grade 4 in 2014.

Regardless of the level of analysis, the baseline equivalence requirement for clusters can be satisfied using individual- or cluster-level means and individual- or cluster-level standard deviations, in any combination, as long as the weighting of the means is consistent with the weighting used in the analysis. The WWC will use individual-level standard deviations when possible. Any required statistical adjustments must be made using data at the same level as those used to assess baseline equivalence.

Additionally, as part of the baseline equivalence requirement for the analytic sample of clusters, the individuals with baseline data must be representative of the clusters contributing to the impact analysis, assessed by comparing the number of individuals contributing baseline data with the number of students in the clusters at the time of the baseline equivalence assessment. This representativeness assessment for baseline data is analogous to the representativeness assessment for follow-up data described in step 5. Baseline characteristics based on administrative data satisfy the representativeness requirement unless review team leadership concludes that relevant individuals are excluded in a nonrandom way from those data. In all other cases, representativeness at baseline is assessed using the same thresholds for attrition in individual-level assignment studies from section II.A. For example, if a school-level assignment study measuring outcomes of grade 4 students in 2015 uses grade 4 students from the same schools in 2014 to assess equivalence, then baseline representativeness would be assessed by comparing the number of grade 4 students enrolled in the schools in 2014 who did not

contribute baseline data to the total number enrolled, both those who did and those who did not contribute data. The timing of the reference sample count needs not be precisely aligned with the collection of the baseline measure. For example, in a school-level assignment study that collected baseline data at the end of a school year and measured outcomes during the following school year, the reference sample might be the administrative school enrollment count taken at some point during the school year in which the baseline data were collected.

WWC review process for step 7 of the review of cluster-level assignment studies

- ▶ If the study satisfies the baseline equivalence requirement for the analytic sample of clusters, including that the baseline data are representative, then the study is eligible to be rated *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study does not satisfy the baseline equivalence requirement for the analytic sample of clusters, then the study is rated *Does Not Meet WWC Group Design Standards*.

Exclusion of sample members in cluster-level assignment studies

Some sample exclusions can be excluded from attrition, nonresponse, and representativeness calculations in cluster RCTs and from representativeness calculations in QEDs. The same criteria about sample loss that is not considered attrition and described for individual-level assignment studies in section II.A apply to cluster-level assignment studies. In particular, when authors analyze outcome data for only a subset of individuals or clusters, the excluded data do not count as attrition if the subsample of the intervention or comparison group was randomly selected or if the subsampling was based on characteristics, like race and gender, that were clearly determined prior to the introduction of the intervention and applied consistently across the intervention and comparison groups.

A cluster RCT is compromised when the study authors do one or more of the following:

1. Include clusters in the analytic sample not subject to random assignment, that is, individuals in the analytic sample who were not subject to random assignment are joiners and are addressed in step 2.
2. Randomly assign clusters to a group with different probabilities but do not use one of the acceptable approaches to account for the different assignment probabilities described in section II.A.
3. Change group membership for an individual or cluster after random assignment.
4. Exclude certain clusters or individuals based on events that occurred after the introduction of the intervention and may have been affected by group status.

When the cluster RCT is compromised for one of these reasons, the study is reviewed using the process for cluster QEDs; that is, the study is not considered a cluster RCT with low cluster-level attrition in steps 1 or 6.

C. Other analytic approaches

Authors of group design studies may use a variety of analytic approaches to measure an intervention's effectiveness or to satisfy the baseline equivalence requirement. Below, the WWC provides guidance on how these different types of analytic approaches, which may be used in individual-level or cluster-level assignment studies, can affect study ratings and reporting of effect sizes and p values. This section provides guidance on the following types of analyses: analyses from propensity score models, analyses in which subjects are observed in multiple time periods, analyses with endogenous covariates, and analyses with missing data. The *WWC Procedures Handbook* provides general information about how the WWC reports study findings.

1. Propensity score analyses

A propensity score is the probability that an observation would appear in the intervention group given a set of measured characteristics. The scores can be used to identify subjects from a pool of potential comparison group members and to match them to intervention group members who have similar characteristics. Alternatively, the scores can be used as weights in a regression analysis designed to make the weighted intervention and comparison groups more similar.

When a study employs propensity-scoring approaches, the WWC will review the study using the same framework as any other QED, requiring that the analytic intervention and comparison groups satisfy the baseline equivalence requirement, including statistical adjustments if necessary. However, for a propensity score analysis to credibly satisfy baseline equivalence for the analytic sample, WWC reviewers must assess the following two key considerations:

- a. *Only exogenous covariates have been used to create the propensity scores.* If potentially endogenous covariates or outcomes are used in the creation of the propensity scores, the scores may ultimately lead to biased impact estimates. See section II.3, Analyses with potentially endogenous covariates, for how the WWC identifies endogenous covariates.
- b. *The analytic approach used to satisfy the baseline equivalence requirement is appropriate.* If the study analysis used propensity score weights, then the baseline means should also be calculated using the same weights. Equivalence must be assessed on the variables specified in the review protocol; it is not sufficient to establish equivalence on the propensity scores. Furthermore, any required statistical adjustments must use the actual specified variables and not only the propensity scores.

Additionally, for the WWC to report the statistical significance of the findings from a propensity score analysis, significance levels must not be artificially inflated due to matching with replacement. Propensity score analyses that use either weighting or matching techniques are acceptable, including matching with replacement. However, if the study used matching with replacement, then reviewers should examine whether the study authors took reasonable precautions in the calculation of standard errors to ensure that the repeated observations of subjects do not contribute to artificially precise estimates. For example, a study might appropriately address this concern by applying a clustering correction to account for the repeated observations.

2. Analyses in which subjects are observed in multiple time periods

This section provides guidance on two types of analyses in which subjects are observed in multiple time periods, sometimes referred to as *repeated measures* analyses: analyses of simple

gain scores and analyses in which the dependent variable includes data from multiple time points. In contrast, the additional considerations for these analyses described below do not apply to analyses in which preintervention measures of the outcome are instead included as covariates in the analytical model. Regardless of the approach used to analyze the repeated measures, the baseline equivalence requirement, if applicable, must still be satisfied on the measures specified in the review protocol. In the case that the baseline difference falls between 0.05 and 0.25 standard deviation, in addition to regression adjustment and analysis of covariance, the WWC considers analyzing simple gain scores, difference-in-differences adjustments, and individual fixed effects as acceptable statistical adjustments, but only when there is evidence that the baseline and outcome measures are strongly related based on the requirements described in section II.A (see table II.3).⁴

Analyses of simple gain scores

Simple gain scores can be calculated by subtracting a pretest from the posttest. Some authors use the resulting difference as the dependent variable in an impact analysis. The analyses of simple gain scores are eligible to meet WWC group design standards. However, to be reported by the WWC, effect sizes from gain score analyses must be based on standard deviations of the outcome measure collected at the follow-up time point without adjustment for the baseline measure; see the gain scores subsection of appendix E of the *WWC Procedures Handbook, Version 4.1*. When the unadjusted standard deviations are not reported but are needed to calculate and report an effect size, the WWC will request the unadjusted posttest standard deviations from study authors. If the WWC cannot calculate an effect size based on acceptable standard deviations, then the study is still eligible to meet WWC group design standards. However, to meet WWC group design standards, a study must report the direction of the impact estimate, for example, whether the difference in means is positive or negative. If the authors do not provide any information about the direction of the impact estimate, then the study is rated *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards.

Analyses in which the dependent variable includes data from multiple time points

In these repeated measures analyses, the analysis includes multiple observations for each student, and the dependent variable includes data from all time points. For example, students are observed in two or more periods, at least one preintervention and one postintervention, and the analysis includes multiple observations for each student, one at each point in time. These include difference-in-differences analyses, comparative interrupted time-series analyses, and most growth curve models. To be eligible for review as a group design study, the study must measure the effect of the intervention by comparing exclusive intervention and comparison groups, meaning that a subject can belong to only a single group at each point in time. Analyses in which the same subject is analyzed as a member of both the intervention and comparison groups at different times are not eligible for review unless there are distinct intervention and comparison groups at each time period after baseline. For example, consider a study of an intervention that is provided to students in group A during period 1 and to students in group B during period 2. Students in groups A and B receive the comparison condition when not receiving the

⁴ The repeated measures analyses discussed in this section—simple gain scores and analyses in which the dependent variable includes data from multiple time points—would rarely use regression adjustment or analysis of covariance to adjust for a preintervention measure of the outcome. However, a repeated measures analysis may use these adjustment approaches to account for other preintervention measures that might be specified in the review protocol.

intervention. If the authors examined the impact of the intervention separately in each period by comparing students who received the intervention in that period with the distinct group of students who did not, then this study would be eligible for review as a group design study. However, if the authors examined the impact of the intervention separately for each group of students by comparing preintervention outcomes with postintervention outcomes, then this study would be ineligible for review as a group design study. When the analysis of such a study does not provide an impact estimate comparing exclusive intervention and comparison groups, the WWC will request means and standard deviations for the exclusive groups from study authors.

The WWC separately reviews impact findings at each point in time included in these analyses. Each impact estimate, or an average of impacts across time periods, is eligible to be rated *Meets WWC Group Design Standards Without Reservations* if the groups were formed in a low-attrition RCT and otherwise is eligible to be rated *Meets WWC Group Design Standards With Reservations*. Returning to the study in the example above, the author of the study might report the weighted or unweighted average of the two impact estimates. The average of the impact estimates at times 1 and 2 will exactly equal the average of the impact estimates for group A and group B, so the method of calculation, regardless of whether it is calculated using point-in-time differences or within-group differences, does not affect the rating. However, if the study is a QED, then the average impact is eligible to be rated *Meets WWC Group Design Standards With Reservations* only if the baseline equivalence requirement is satisfied separately for the two time points. Although the groups may have been equivalent at the start of period 1, exposure to the intervention for subjects in group A during period 1 might lead to differences in the groups at the start of period 2. Alternatively, if subjects were randomly assigned to conditions in period 1, then the WWC will review the period 1 finding as an RCT, while the WWC will review the period 2 impact estimate as a QED that must satisfy the baseline equivalence requirement because subjects in group A were exposed to the intervention.

Growth curve analyses do not typically provide point-in-time impact estimates. However, the WWC will request the data needed from authors to calculate effect sizes—and baseline equivalence, if required—at each point in time. If the WWC cannot calculate an effect size, then the study is still eligible to meet WWC group design standards. However, to meet WWC group design standards, a study must report the direction of the impact estimate at the specific point in time. If the authors do not provide any information about the direction of the impact estimate, then the study is rated *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards.

To be eligible to meet WWC group design standards, the analysis must adequately account for the time periods associated with the intervention and preintervention conditions. In a difference-in-differences analysis, which includes just two time periods—preintervention and postintervention, this means including indicators for the intervention condition, the time period associated with the intervention, and an interaction between these two indicators. In such an analysis, the coefficient on the interaction term provides the difference-in-differences estimate of the impact of the intervention, and the *p* value of this estimate is used to assess the statistical significance of the impact. A mixed design analysis of variance with one between-groups factor distinguishing the intervention and comparison groups and at least one within-groups factor distinguishing time period typically satisfies this requirement. Studies may sometimes refer to this as a repeated measures analysis of variance. This requirement can also be satisfied by an

ordinary least squares (OLS) analysis that includes the intervention, time period, and interaction indicators as independent variables.

A study that instead reports the coefficient on the intervention indicator and excludes the interaction provides a biased estimate of the effect of the intervention because doing so measures the average difference in the outcome between the intervention and comparison groups across both the preintervention and postintervention periods. Such an analysis does not provide a credible estimate of the effectiveness of the intervention, and if the authors do not provide the WWC with findings from a credible analysis, then the study will be rated *Does Not Meet WWC Group Design Standards*.

Analyses with more than two time periods, including most comparative interrupted time-series and growth curve analyses, must also account for the preintervention and postintervention periods, including an interaction with the intervention indicator, but they also can also account for additional time periods. Adjusted or unadjusted means and unadjusted standard deviations of the outcome at each postintervention time point can also be used to satisfy this requirement. In analyses that include multiple periods of preintervention data, baseline equivalence must be assessed using data from a single period so that the intervention period can be defined as the time from the baseline assessment to follow-up. The WWC will use the preintervention time point closest to the introduction of the intervention to assess baseline equivalence, when possible.

3. Analyses with potentially endogenous covariates

Reviewers should examine model specifications and descriptions of analytic procedures to ensure that the estimates of intervention effects from study-reported analyses are credible, given the proposed analytic procedure. In some impact evaluations, researchers will estimate regression models—for example, ordinary least squares or hierarchical linear modeling—in which the outcome of interest is regressed on an indicator for the intervention and a series of covariates.

Reviewers should determine whether a study includes covariates in its impact analyses that were assessed or obtained after baseline. If such variables are included and were *potentially influenced by group status*, then the impact analysis will produce a biased test of the effect of the intervention. In contrast, if a covariate is obtained after baseline and is unlikely to have been influenced by group status or is considered time invariant—for example, demographics such as sex and race—then there is no concern that the variable is an endogenous covariate influenced by the intervention.

For example, a study that examines the impact of an intervention on student achievement outcomes may collect data on student attendance during the intervention or the quality of teacher–student interactions. These variables associated with intervention dose, quality, or fidelity may have been affected by the intervention. If the impact analysis includes either student attendance during the intervention or the quality of teacher–student interactions as covariates, then the correlation between the intervention indicator and these variables will produce bias in the impact estimate. Therefore, the WWC cannot use the results of the regression model as a credible source of information about an intervention’s effects.

A measure assessed shortly after the start of the intervention is not considered to be a potentially endogenous covariate (Schochet, 2008). A measure assessed later, after the

intervention may have plausibly affected the measure in the judgement of the review team leadership and content expert, is a potentially endogenous covariate. However, if the potentially endogenous measure is used to satisfy the baseline equivalence requirement described in step 3 in section II.A, then the WWC will note in its reporting that the study measures the effect of the portion of the intervention that occurred after the measure was assessed and until the time of the follow-up assessment. Even though the baseline measure may have been influenced by the intervention, it can be used to satisfy the baseline equivalence requirement. It is not necessary to include the same reporting note for a baseline measure assessed shortly after the start of the intervention that was included as a covariate in the analysis, but it is not used to satisfy the baseline equivalence requirement.

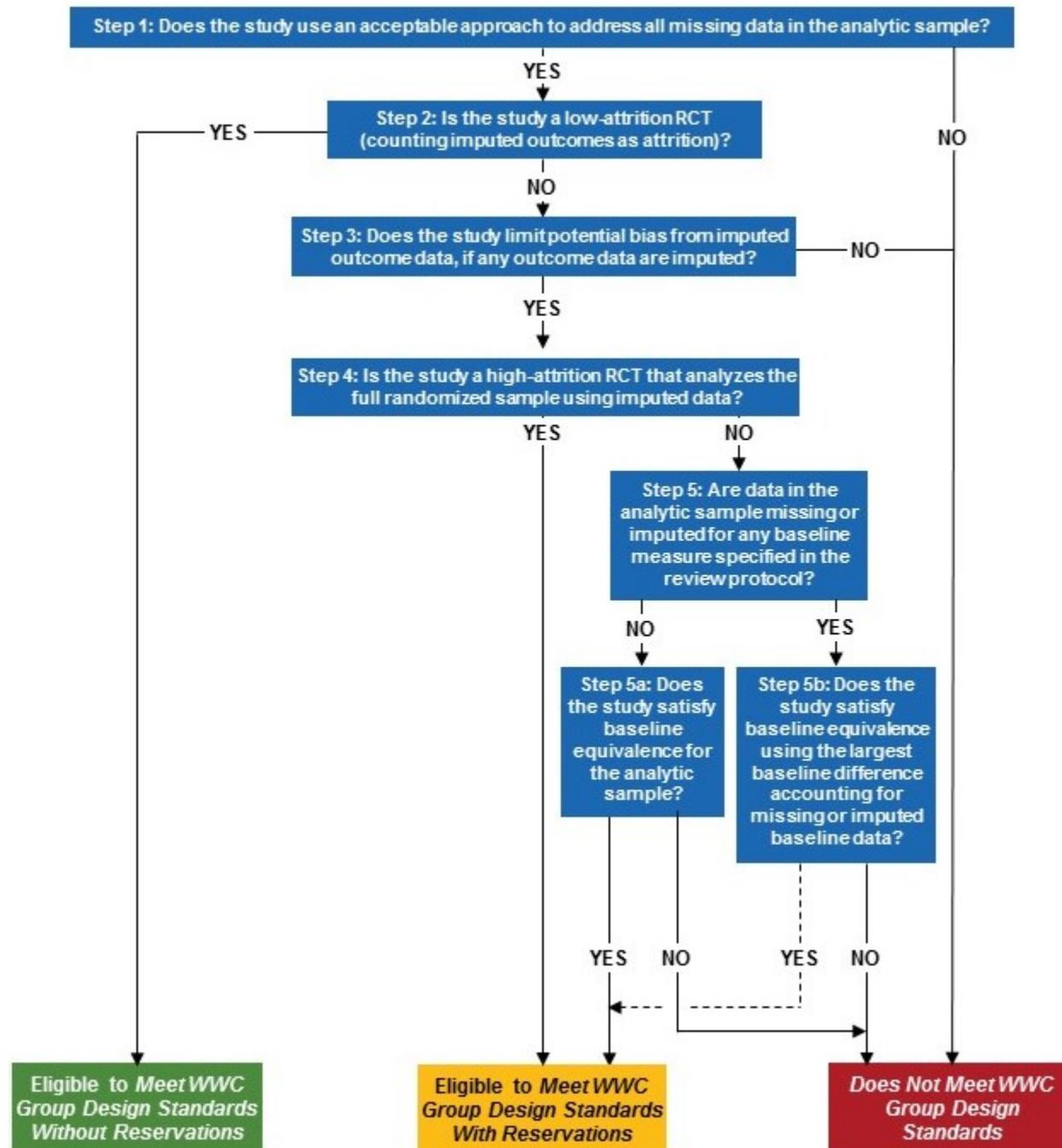
When one or more potentially endogenous covariates are included in the analysis, the WWC can either use alternative model specifications reported in the study that do not include these endogenous covariates or request unadjusted means—or adjusted means based on only the nonendogenous covariates—and unadjusted standard deviations from the authors. However, to be eligible to meet WWC group design standards, a study must report the direction of the impact estimate from a credible analysis. If the authors do not provide any information about the direction of the impact estimate from a credible analysis, then the study is rated *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards.

4. Analyses with missing data

Despite the best efforts of researchers, sometimes it is not possible to collect data for all subjects in a study sample. Authors might use a variety of analytical approaches to address missing data for baseline or outcome measures. For example, a study might focus on the analytic sample of subjects for which all data were collected, or the authors may impute values for the missing data so that more subjects can be included in the analysis. The review process for a study with missing data depends on the study design, the method used to address the missing data, and whether the study has missing baseline data, outcome data, or both.

The steps in the review process for studies with missing data are outlined in figure II.5. Steps 1 and 2 must be performed for any study with missing data, steps 3 and 4 relate to studies with imputed outcome data in the analytic sample, and step 5 relates to studies with imputed or missing baseline data in the analytic sample. We describe each of these steps in detail next.

Figure II.5. Study ratings for randomized controlled trials and quasi-experimental designs with missing outcome or baseline data



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in chapter V, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Step 1. Does the study use an acceptable approach to address all missing data in the analytic sample?

The first step in the review process for studies with missing data is to determine whether any imputed data used in the analysis were generated using an acceptable imputation method. To be eligible to be rated *Meets WWC Group Design Standards With or Without Reservations*, an analysis must use one of the methods described in table II.6 to address the missing data. This requirement applies to all data used in the analysis, whether for an outcome measure or a baseline measure. More specifically, the requirement applies both to baseline measures specified in the review protocol as required for assessing baseline equivalence and those not specified. Analyses that include any imputed outcome or baseline data based on other approaches not listed in table II.6 are rated *Does Not Meet WWC Group Design Standards*.

When an analysis uses one or more of these methods and satisfies all other requirements to receive a rating of *Meets WWC Group Design Standards With or Without Reservations*, the WWC will report findings, including effect sizes, according to the general approach to WWC reporting outlined in the *WWC Procedures Handbook*. However, the WWC will not report statistical significance for methods that do not provide accurate standard error estimates. For some other methods, the WWC will report statistical significance provided certain requirements are met, as described in the last column in table II.6.

All but one of the acceptable approaches in table II.6 can provide unbiased estimates of the effectiveness of an intervention based on the assumption that the missing data do not depend on unmeasured factors. The exception is complete case analysis, which requires a more restrictive assumption that the missing data also do not depend on measured factors. Because of this, many researchers have recommended against using complete case analysis to address missing data (for example, Little et al., 2012; Peugh & Enders, 2004). Nevertheless, the WWC considers complete case analysis to be an acceptable approach for addressing missing data because possible bias due to measured factors can be assessed through the attrition standard and WWC's baseline equivalence requirement; see steps 2 and 3 of section II.A.

In addition, Jones (1996) and Allison (2002) raised concerns about using the approach in the last row table II.6, imputation to a constant combined with including a missing data indicator, outside of RCTs. Consequently, the WWC considers this approach acceptable for any baseline data in RCTs regardless of their sample attrition. However, in a QED or compromised RCT, the approach is acceptable only when applied to baseline measures not specified in the review protocol as required for assessing baseline equivalence.

To obtain appropriate estimates of statistical significance in cluster-level assignment studies that analyze individual-level data, approaches to address missing outcome data must account for the correlation of outcomes within clusters. This can be done using standard approaches in complete case analyses. However, as noted in the last column of table II.6, for the WWC to confirm statistical significance in a study with cluster-level assignment that uses regression imputation, maximum likelihood, or nonresponse weights to address missing outcome data, and analyzes individual-level data, the study must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness. In analyses using these three approaches that do not include an acceptable adjustment, the WWC will not apply its adjustment for clustering, as described in the *WWC Procedures Handbook*, because it may not be accurate.

for analyses using these methods. The WWC does not currently have a recommended method of calculating standard errors in these analyses of cluster-level assignment studies, and the burden for demonstrating that the approach is appropriate rests with the study authors.

WWC reviewers do not receive training on the approaches listed in table II.6, but understanding their application can require specialized knowledge and training. Reviewers should bring questions about whether a study appropriately applied any of these methods to the review team leadership.

Finally, if a study uses an approach not listed in table II.6 that is supported with a citation to a peer-reviewed journal article or textbook that describes the procedure and demonstrates that it can produce unbiased estimates under an assumption that the missing data are unrelated to unmeasured factors, the WWC may consider it an acceptable approach after review by experts. If so, the WWC will release guidance that updates the list of acceptable approaches.

Table II.6. Acceptable approaches for addressing missing baseline or outcome data

Approach	Description	WWC requirements	Statistical significance
Complete case analysis	Exclusion of observations with missing outcome and/or baseline data from the analysis.	None.	The WWC has no additional requirements for reporting statistical significance from analyses that use this method.
Regression imputation	A regression model to predict imputed values for the missing data. This includes estimating imputed values from a single regression model, and multiple imputation, which involves generating multiple datasets that contain imputed values for missing data through the repeated application of an imputation algorithm, such as chained equations.	The imputation regression model must: <ol style="list-style-type: none"> Be conducted separately for the intervention and comparison groups or include an indicator variable for intervention status, Include all of the covariates that are used for statistical adjustment in the impact estimation model, and Include the outcome when imputing missing baseline data. 	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or multiple imputation. For multiple imputation, the statistical significance calculation must: <ol style="list-style-type: none"> Be based on at least five sets of imputations, and Account for (1) the within-imputation variance component, (2) the between-imputation variance component, and (3) the number of imputations. Most established multiple imputation routines satisfy this requirement. Additionally, a cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Maximum likelihood	An iterative routine to estimate model parameters and impute values for the missing data. Some examples are the expectation-maximization algorithm and full	The procedure must use a standard statistical package or be supported with a citation to a peer-reviewed methodological journal article or textbook.	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or estimates based on the information matrix. Additionally, a cluster-level assignment study with missing outcome data, analyzed using

Approach	Description	WWC requirements	Statistical significance
Nonresponse weights	information maximum likelihood.		individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Replacing missing data with a constant combined with including a missing data indicator	Use of weights based on estimated probabilities of having a nonmissing outcome, yielding greater weight for subjects with a higher probability of having missing outcome data. For example, the probabilities may be estimated from a logit or probit model.	<p>Acceptable only for missing outcome data, not for missing baseline data.</p> <p>The estimated probabilities used to construct the weights must:</p> <ol style="list-style-type: none"> Be estimated separately for the intervention and comparison groups or include an indicator variable for intervention status, and Include all baseline measures that are specified in the review protocol as required for baseline equivalence within the outcome domain. Including additional covariates is acceptable but not required because doing so may lead to less precise impact estimates without providing a substantial reduction in bias. 	<p>The analysis must properly account for the stratified sampling associated with the weights (as discussed in Wooldridge, 2002, p. 594).</p> <p>Additionally, a cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.</p>

Note:

Requirements in this table are based on recommendations in several sources, including Allison (2002), Azur, Stuart, Frangakis, and Leaf (2011); Little and Rubin (2002); Puma, Olsen, Bell, and Price (2009); Rubin (1987); Schafer (1999); and Wooldridge (2002).

WWC review process for step 1 of the review of studies with missing data

- If the study uses an acceptable approach to address all missing data in the analytic sample, then continue to step 2.
- If the study does not use an acceptable approach to address all missing data in the analytic sample, then the study is rated *Does Not Meet WWC Group Design Standards*.

Step 2. Is the study a low-attrition randomized controlled trial (counting imputed outcomes as attrition)?

The second step in the review process for studies with missing data is to determine whether the study is a low-attrition RCT as described in step 2 of section II.A. When calculating overall and differential attrition rates, sample members with imputed outcome data are counted as missing because both missing and imputed data represent a potential threat of bias. The use of imputed data can mitigate that bias if the missing data do not depend on unmeasured factors, but otherwise may not. When attrition is low, the WWC will ignore the potential bias from imputed

data because the amount of missing or imputed data is unlikely to lead to bias that exceeds the WWC's tolerable level of potential bias. A low-attrition RCT is eligible to be rated *Meets WWC Group Design Standards Without Reservations* as long as the study used an acceptable method to address missing data.

WWC review process for step 2 of the review of studies with missing data

- ▶ If the study is a low-attrition RCT, then the study is eligible to receive the rating *Meets WWC Group Design Standards Without Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study is a QED, high-attrition RCT, or compromised RCT, then continue to step 3 of the review process for studies with missing data.

Step 3. Does the study limit potential bias from imputed outcome data, if any outcome data are imputed?

Imputed outcome data can affect the rating of a QED, high-attrition RCT, or compromised RCT in two ways. The first of these is addressed in step 3. To be eligible for a rating of *Meets WWC Group Design Standards With Reservations*, QEDs, high-attrition RCTs, and compromised RCTs with imputed outcome data in the analytic sample must satisfy an additional requirement designed to limit potential bias from using imputed outcome data instead of actual outcome data.

The imputation methods the WWC considers acceptable are based on an assumption that the missing data depend on measured factors, not unmeasured factors. If that assumption does not hold, then impact estimates may be biased. Therefore, group design studies besides low-attrition RCTs that use acceptable approaches to impute outcome data must demonstrate that they limit the potential bias from using imputed data to measure impacts to less than 0.05 standard deviation as described in this step.

An analysis of a sample with imputed outcome data can produce biased estimates of the effect of the intervention if the subjects with observed data differ from the subjects with missing data, and some of the differences are unmeasured. In this case, if outcomes could be obtained for all sample members, then the average for subjects in the intervention or comparison condition with observed outcome data would differ from the average for subjects whose outcome data were not observed. Comparing the differences in these means for the intervention and comparison groups, if known, would indicate the magnitude of possible bias, but because the missing outcomes are not observed, the WWC instead assesses the bias using baseline data.

The WWC estimates the potential bias from missing outcome data due to unmeasured factors by comparing means of the baseline measure specified in the review protocol as required for assessing baseline equivalence, separately for the intervention and comparison groups, for two samples: the complete analytic sample and the analytic sample restricted to cases with observed outcome data. A smaller difference in these two means within one or both conditions lowers the likelihood that the missing data are related to factors that could lead to bias in the impact estimate.

To translate the intervention and comparison group differences in baseline means into an estimate of bias in the outcome effect size, the WWC uses the pooled standard deviation of the

baseline measure and the correlation between the baseline and outcome measure. Appendix A provides the formulas the WWC uses to estimate the potential bias (equations A.5.0–A.5.2). Appendix A also describes the approach used when a review protocol specifies that baseline equivalence must be assessed on multiple baseline measures. The formulas used to assess the bias also differ depending on whether the baseline measure is observed for all subjects in the analytic sample (equations A.10.0–A.10.2).

- **When the baseline measure is observed for all subjects in the analytic sample**, the WWC requires the following data from the authors: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups—these are the same data used to assess baseline equivalence; (b) the means of the baseline measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups; and (c) the correlation between the baseline and the outcome measures. The correlation can be estimated on a sample other than the analytic sample, such as the complete case sample, or from data from outside the study if a content expert judges the settings to be similar. However, the correlation must not be estimated using imputed data.
- **When the baseline measure is imputed or missing for some subjects in the analytic sample**, in addition to (c), the following data are required: (d) the means of the baseline measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups; (e) the means of the baseline measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (f) the standard deviations of the baseline measure for either the sample of subjects in the analytic sample with observed baseline data or the sample with observed baseline and outcome data; and (g) the number of subjects with observed baseline data in the analytic sample by condition.

If these data are not reported in the study, then the WWC will request them from the authors.

There are two special considerations for applying the requirement in step 3 when an analysis uses nonresponse weights or complete case analysis:

- An analysis that uses nonresponse weights to address missing outcome data must also satisfy the requirement to limit the potential bias from using imputed data. For these analyses, separately for the intervention and comparison groups, the WWC compares a different pair of means of the baseline measure. Instead of the complete analytic sample, which for a nonresponse weighted analysis would be restricted to cases with observed outcome data, the WWC uses the sample used to estimate the weights, including cases with missing outcome data. The second mean remains the sample with observed outcome data.
- A complete case analysis that addresses missing data by excluding cases with missing outcome data, rather than imputing it, does not need to satisfy this requirement. The exclusion of complete case analyses from this requirement is not intended to imply that complete case analyses are believed to be a stronger approach for addressing missing data. Rather, the WWC's approach recognizes that the attrition standard and baseline equivalence requirement can limit bias in complete case analyses because the missing data affect the analytic sample.

WWC review process for step 3 of the review of studies with missing data

- ▶ If the study limits potential bias from imputed outcome data, as assessed using the formulas in appendix B, or the analytic sample contains no imputed outcome data, then continue to step 4 of the review process for studies with missing data.
- ▶ If the study does not limit potential bias from unmeasured factors, then the study is rated *Does Not Meet WWC Group Design Standards*.

Step 4. Is the study a high-attrition RCT that analyzes the full randomized sample using imputed data?

The fourth step in the review process for missing outcome data addresses a second way imputed outcome data can affect the rating of a study. When study authors analyze a high-attrition RCT by imputing outcome data so that they analyze the full sample that was randomized to conditions, the study does not need to satisfy the baseline equivalence requirement to be eligible to receive the rating *Meets WWC Group Design Standards With Reservations*.

In general, the WWC requires that high-attrition RCTs satisfy the baseline equivalence requirement because of a risk of bias from compositional differences between the remaining intervention and comparison group members. However, some high-attrition RCTs impute all missing outcome data and analyze the original randomized sample. These high-attrition RCTs do not need to satisfy the baseline equivalence requirement because of a presumption that intervention and comparison groups that result from random assignment are unlikely to have substantive compositional differences. Imputing missing outcome data and analyzing the full randomized sample preserves the integrity of the originally randomized groups. Although compositional differences are not considered a threat to bias, like other high-attrition RCTs, these studies are eligible to be rated only *Meets WWC Group Design Standards With Reservations*. These studies are not eligible for the highest rating because of the risk of bias from imputing a larger amount of missing outcome data compared with a low-attrition RCT.

All QEDs, high-attrition RCTs that do not analyze the original randomized sample, and compromised RCTs must satisfy the baseline equivalence requirement (step 5 in figure II.5).

WWC review process for step 4 of the review of studies with missing data

- ▶ If the study is a high-attrition RCT that analyzes the original randomized sample, then the study is eligible to receive the rating *Meets WWC Group Design Standards With Reservations* and does not need to satisfy the baseline equivalence requirement. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study is a QED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, then the study must satisfy the baseline equivalence requirement to be eligible to receive the rating *Meets WWC Group Design Standards With Reservations*. Continue to step 5 of the review process for studies with missing data.

Step 5. Are data in the analytic sample missing or imputed for any baseline measure specified in the review protocol?

QEDs, high-attrition RCTs that do not impute data to analyze the full randomized sample, and compromised RCTs must satisfy the baseline equivalence requirement to be eligible to be rated *Meets WWC Group Design Standards With Reservations*. However, it is not possible for the WWC to assess baseline equivalence on the full analytic sample using actual data when some

data are missing or imputed for a measure that is specified in the review protocol as required for assessing baseline equivalence.

WWC review process for step 5 of the review of studies with missing data

- ▶ If the study is a QED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, and the analytic sample does not include missing or imputed data for any baseline measure specified in the review protocol, then continue to step 5a of the review process for studies with missing data.
- ▶ If the study is a QED, high-attrition RCT that does not analyze the original randomized sample, or a compromised RCT, and the analytic sample includes some missing or imputed data for a baseline measure specified in the review protocol, then continue to step 5b of the review process for studies with missing data.

Step 5a. Does the study satisfy baseline equivalence for the analytic sample?

If all of the missing or imputed baseline data in the analytic sample are for baseline measures not specified in the review protocol as required for satisfying baseline equivalence in the outcome domain, or no baseline data are missing or imputed, then baseline equivalence can be assessed using the usual approach described in step 3 of section II.A. A study that satisfies the baseline equivalence requirement using actual data for the analytic sample is eligible to be rated *Meets WWC Group Design Standards With Reservations*.

An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.

WWC review process for step 5a of the review of studies with missing data

- ▶ If the study satisfies the baseline equivalence requirement using actual baseline data, the study is eligible to receive the rating *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- ▶ If the study does not satisfy the baseline equivalence requirement using actual baseline data, the study is rated *Does Not Meet WWC Group Design Standards*.

Step 5b. Does the study satisfy baseline equivalence using the largest baseline difference accounting for missing or imputed baseline data?

If some data are missing or imputed for a baseline measure that is specified in the review protocol as required for satisfying baseline equivalence in the outcome domain, then the WWC uses a different process to assess baseline equivalence. In this case, the WWC estimates how large the baseline difference might be under different assumptions about how the missing data are related to measured or unmeasured factors. The largest of these estimates in absolute value is used as the baseline difference for the study.

Just as for studies with complete baseline data, a study with missing or imputed data for a required baseline measure is eligible to be rated *Meets WWC Group Design Standards With Reservations* if the largest estimated standardized baseline difference does not exceed 0.25 standard deviation when the analysis includes an acceptable adjustment for the baseline measure, or 0.05 standard deviation otherwise. A study that satisfies this alternative baseline equivalence requirement is eligible to be rated *Meets WWC Group Design Standards With Reservations*.

The WWC's approach to estimating the baseline difference in studies with missing or imputed baseline data is similar to the approach used to estimate bias from using imputed outcome data, described above. Instead of comparing means of the baseline measure, the WWC compares means of the outcome measure, separately for the intervention and comparison groups, for two samples: the analytic sample and the analytic sample restricted to cases with observed baseline data. A larger absolute difference in these means within a group indicates that the data may be missing in a way that is related to unmeasured sample characteristics, and the measured impact of the intervention may be biased.

To translate the intervention and comparison group differences in outcome means into an estimate of a baseline effect size, the WWC uses the pooled standard deviation of the outcome measure and the correlation between the baseline and outcome measure. Appendix B provides the formulas the WWC uses to estimate the baseline effect size (equations B.5.0–B.5.3, B.7.0–B.7.3, B.11.0–B.11.3, and B.13.0–B.13.3). When a review protocol specifies that baseline equivalence must be assessed on multiple baseline measures, the formulas in appendix B must be applied to each required baseline measure. The formulas used to estimate the baseline difference vary based on two factors: whether the outcome measure is observed for all subjects in the analytic sample and whether the outcome data are missing or imputed.

- **When the outcome measure is observed for all subjects in the analytic sample**, the WWC requires the following data from the authors: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups; (b) the means of the outcome measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups; (c) the correlation between the baseline and the outcome measures; and (d) an estimate of the baseline difference based on study data. As noted in step 3 of the section on imputed outcome data, the correlation can be estimated on a sample other than the analytic sample but must not be estimated using imputed data. If the authors did not impute the baseline data, then the WWC will use baseline means and standard deviations to measure the baseline difference for the portion of the analytic sample with observed baseline data. However, if the study did impute baseline data, then the WWC will include the imputed data when calculating the means but will use standard deviations based only on the observed data.
- **When the outcome measure is imputed for some subjects in the analytic sample**, in addition to (c) and (d), the following data are required: (e) the means of the outcome measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups; (f) the means of the outcome measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (g) the standard deviations of the outcome measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data; and (h) the number of subjects with observed outcome data in the analytic sample by condition.

If these data are not reported in the study, then the WWC will request them from the authors.

The two special considerations for applying the requirement in step 5b when an analysis uses nonresponse weights or complete case analysis are as follows:

- An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.
- Because no baseline data are missing or imputed, a complete case analysis that excludes cases with missing baseline data must satisfy the baseline equivalence requirement using the observed data for the analytic sample, as described above in step 3 of section II.A, rather than using the formulas in appendix B. In other words, the complete case analysis must satisfy baseline equivalence using step 5a and not step 5b.

WWC review process for step 5b of the review of studies with missing data

- If the study satisfies the baseline equivalence requirement using the largest baseline difference (estimated according to the formulas in appendix B) accounting for the missing or imputed data, the study is eligible to receive the rating *Meets WWC Group Design Standards With Reservations*. To receive this rating, the study must also satisfy the requirements in chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.
- If the study does not satisfy the baseline equivalence requirement using the largest baseline difference accounting for the missing or imputed data, then the study is rated *Does Not Meet WWC Group Design Standards*.

D. Complier average causal effects

In RCTs, subjects are randomly assigned to groups that differ in access to an intervention. However, subjects do not always comply with their assigned conditions. In the assigned intervention group—the group whose assignment makes them eligible for the intervention—some subjects might choose not to receive intervention services. In the assigned comparison group—the group whose assignment makes them ineligible for the intervention—some subjects might nevertheless receive the intervention.

In the presence of noncompliance, RCT studies have typically estimated either or both of two impacts. First, to estimate the effect of being assigned to the intervention, known as the ITT effect, the mean difference in outcomes between the *entire* assigned intervention group and the entire assigned comparison group is calculated.

Second, to estimate the effects of actually receiving the intervention, one common approach is to estimate the complier average causal effect (CACE).⁵ The CACE is the average effect of taking up the intervention among compliers—those who would take up the intervention if assigned to the intervention group and who would not take up the intervention if assigned to the comparison group.

The CACE cannot be estimated with a subgroup analysis because compliers cannot be fully distinguished from other sample members. In particular, among sample members assigned to the intervention group, compliers cannot be distinguished from *always-takers*—those who would always take up the intervention, regardless of their randomly assigned status—because both groups take up the intervention. Among sample members assigned to the comparison group, compliers cannot be distinguished from *never-takers*—those who would never take up the

⁵ In some disciplines, the CACE is also referred to as the local average treatment effect. Seminal papers by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) provide a formal discussion of how the CACE can be identified and estimated.

intervention, regardless of their randomly assigned status—because neither group takes up the intervention.

Instead, the CACE is typically estimated with an instrumental variable (IV) estimator, which uses only the variation in take-up that is induced by the random assignment process to estimate the impacts of taking up the intervention on outcomes. An IV estimator starts from an assumption, known as the exclusion restriction, that neither the outcomes of always-takers nor the outcomes of never-takers differ between the intervention and comparison groups (because assignment to those groups cannot influence their take-up status). Any difference between the intervention and comparison groups must therefore be attributable to compliers. Likewise, the difference in take-up rates between the two groups reveals the fraction of study sample members who are compliers. Conceptually and, in certain scenarios, mathematically, an IV estimator, therefore, estimates the effect of the intervention on compliers by dividing the difference in outcomes between the intervention and comparison groups by the difference in take-up rates. As discussed later, conventional statistical tests based on IV estimators perform well only if sample members' randomly assigned status has a strong association with take-up.

This section is intended to specify the scenarios under which CACE estimates from RCTs are eligible for review and subsequently eligible to be rated *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*.

1. Criteria for whether RCT studies are eligible for review under CACE standards

To be eligible for review, a CACE estimate from an RCT must meet several technical criteria. To specify these technical criteria, it is necessary to define some key terms, as discussed next.

Key terms

We refer to the following commonly accepted terms from the econometric literature on instrumental variables:

- *Endogenous independent variable*: The variable whose impact on outcomes is the impact of interest. In this context, the endogenous independent variable is a binary indicator for taking up the intervention. It is *endogenous* because its variation could be affected by subjects' decisions. A particularly uninterested member of the intervention group might elect not to participate, and the unobserved factors underlying the decision might also be correlated with outcomes, inducing a correlation between take-up and outcomes that is not reflective of a causal effect of the intervention itself.
- *Structural equation*: An equation that models the outcome as a function of the endogenous independent variable and possibly other covariates. In this context, estimation of the structural equation produces an estimate of the CACE—the impact of intervention take-up on outcomes.
- *Instrumental variables*: Variables that induce variation in the endogenous independent variable but are assumed to be uncorrelated with other factors influencing the outcome variable. By definition, instrumental variables are excluded from the structural equation. In this context, the instrumental variables are binary indicators for the group to which subjects were randomly assigned.

- *First-stage equation:* An equation that models the endogenous independent variable as a function of the instrumental variables and possibly other covariates. In this context, the first-stage equation is modeling the extent to which take-up is influenced by randomly assigned group status. Assigned group status should influence take-up because sample members assigned to the intervention group are supposed to receive the intervention and those assigned to the comparison group are not.

Technical eligibility criteria

To be eligible for review under the CACE guidance, a CACE estimate from an RCT must be based on statistical methods that meet all of the conditions listed next.

The endogenous independent variable must be a binary indicator for taking up any portion of the intervention. The WWC does not yet have standards for evaluating studies that estimate the relationship between an outcome and a continuous measure of intervention dosage, so the endogenous independent variable must be binary. Moreover, because it is possible that any positive dosage of the intervention could affect outcomes, the endogenous independent variable must distinguish sample members who took up any portion of the intervention from those who did not.

Each structural equation estimated by the study must have exactly one endogenous independent variable. With multiple endogenous independent variables, criteria for evaluating instrument strength (see Stock & Yogo, 2005) would require matrix algebraic quantities that are rarely reported in education evaluations.⁶

The instrumental variables must be binary indicators for the intervention and comparison groups to which subjects are randomly assigned. If random assignment forms two assignment groups—one assigned intervention group and one assigned comparison group—then there will be one instrumental variable, a binary indicator that distinguishes the groups.

In some cases, a CACE estimate may use multiple instrumental variables that induce variation in a single endogenous independent variable. For example, if random assignment is conducted separately in several sites, then a study could interact the intervention assignment indicator with site indicators, and then use both the intervention assignment indicator and the interaction terms as instruments. The site indicators would serve as covariates in both the first-stage and structural equations. The use of these multiple instruments allows the first-stage equation to model variation across sites in the extent to which assignment to the intervention group influences take-up.⁷ Another example in which multiple instrumental variables may be warranted is when there are three or more groups—for instance, a group with highest assigned

⁶ With multiple endogenous independent variables, evaluating instrument strength would require calculating the Cragg–Donald statistic, which is the minimum eigenvalue from the matrix analog of the first-stage F statistic (Cragg & Donald, 1993; Sanderson & Windmeijer, 2016; Stock & Yogo, 2005). Many applied researchers would find it challenging to calculate this statistic unless they had access to specific software that performs this calculation, for instance, the “ivregress” command in Stata. Moreover, if a study did not report this statistic, then the WWC would not be able to calculate it without the individual-level data used for the evaluation.

⁷ A multisite CACE estimate does not have to use site-specific intervention assignment indicators; a single intervention assignment indicator can serve as the sole instrumental variable, in which case the study is choosing not to model differences across sites in the effects of intervention assignment on take-up.

priority for receiving the intervention, a group with lower assigned priority, and an assigned comparison group that cannot receive the intervention—to which each subject could be randomly assigned. In this scenario, the instrumental variables are binary indicators for all but one of the assignment groups.⁸

The sets of baseline covariates—*independent variables other than the endogenous independent variable and instrumental variables*—must be identical in the structural equation and first-stage equation. If baseline covariates are included in the analysis, then the structural equation and first-stage equation must contain identical sets of baseline covariates, or else the study will violate either an eligibility criterion specified above or technical conditions needed for model estimation. In particular, if a baseline covariate from the first-stage equation is not included in the structural equation, then it is effectively serving as an instrumental variable that is not among the types of eligible instruments. If a baseline covariate from the structural equation is not included in the first-stage equation, then the model will lack enough sources of variation to estimate all of the coefficients in the structural equation—a scenario known as under-identification.

The study must estimate the CACE using two-stage least squares (2SLS) or a method that produces the same estimate as 2SLS. In 2SLS, the estimated impact of take-up on outcomes is equivalent to that produced by the following two stages. First, the first-stage equation is estimated with OLS, and predicted values of take-up are obtained from these estimates. Second, the endogenous take-up variable is replaced by its predicted values in the structural equation, which is then estimated by OLS. From this second stage, the estimated coefficient on the predicted take-up variable is equivalent to the 2SLS estimate of the CACE, and the standard error of the coefficient must be adjusted to account for the first-stage prediction, as discussed next.

When there is only one instrument, the 2SLS estimate is the same as a ratio in which the numerator is the ITT estimate and the denominator is the estimated effect of intervention assignment on take-up from the first-stage equation. This ratio is similar to, but more general than, the Bloom (1984) adjustment. The Bloom (1984) estimator is the ITT estimate divided by the take-up rate in the intervention group. It is equivalent to the 2SLS estimator when there is no take-up in the comparison group and no baseline covariates are included in the analysis. When these two conditions hold, these standards can be applied to studies that use the Bloom adjustment.⁹

Although 2SLS is the most widely used approach to CACE estimation, other methods exist. Alternative methods include limited information maximum likelihood (Anderson & Rubin, 1949), generalized method of moments (Hansen, 1982), and missing-data methods based on Bayesian procedures or the expectation-maximization algorithm (Imbens & Rubin, 1997a). Because these methods have not been used frequently in education evaluations, we have not

⁸ In all of these examples, there is still only a single take-up variable, and thus, the study still estimates a single average impact of take-up on outcomes.

⁹ When members of the assigned comparison group take up the intervention, the Bloom adjustment is not applicable. When the structural equation has baseline covariates, the Bloom adjustment implicitly excludes those covariates from the first-stage equation, leading to underidentification.

proposed standards that apply to these methods, and analyses using these methods are ineligible for review.

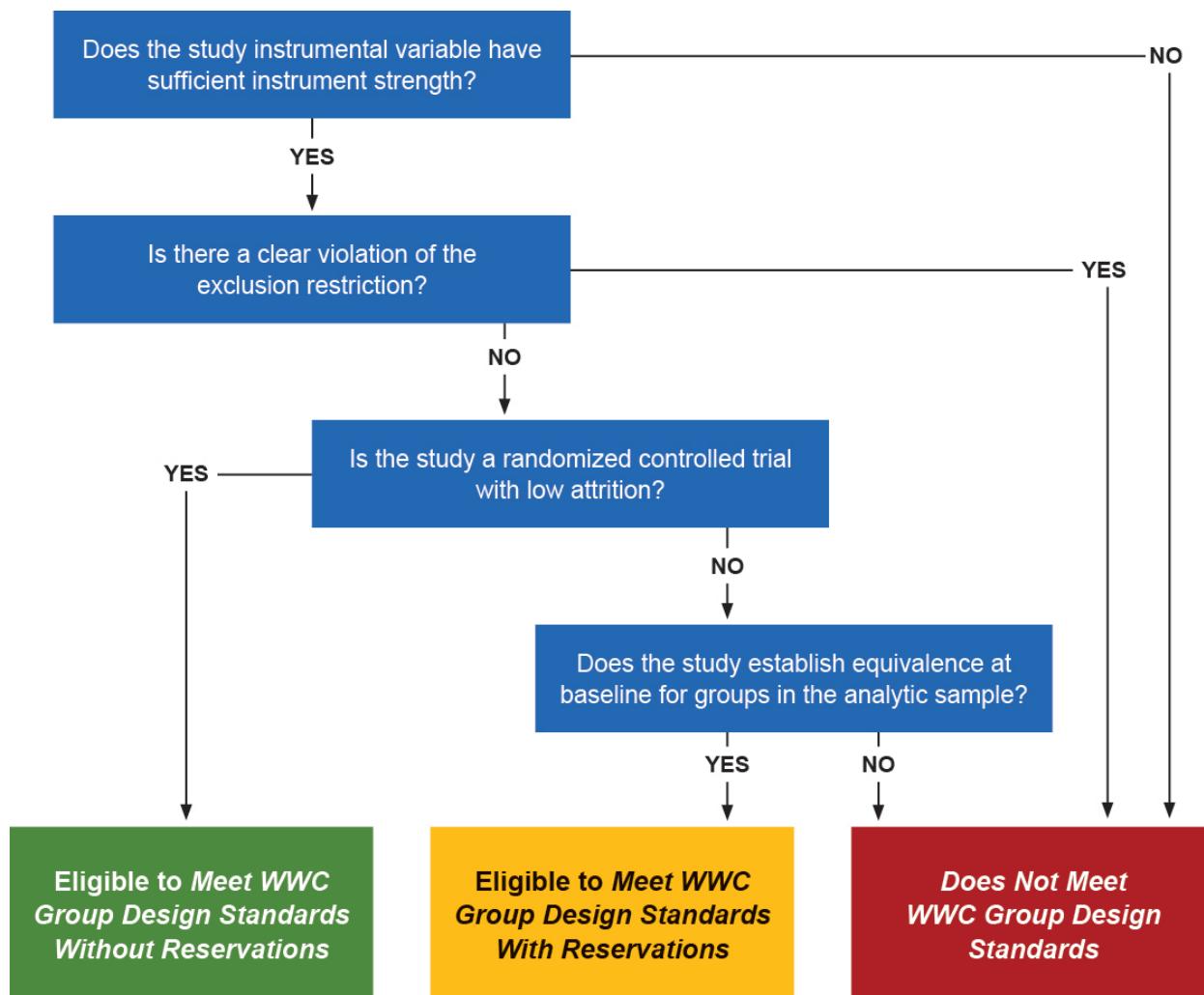
2. Overview of the process for rating CACE estimates

A CACE estimate from an RCT is evaluated on a different set of criteria depending on whether the RCT has low or high attrition, as follows:

- **A CACE estimate from an RCT with low attrition** is rated *Meets WWC Group Design Standards Without Reservations* if it satisfies two conditions: no clear violations of the exclusion restriction and sufficient instrument strength.¹⁰ It is rated *Does Not Meet WWC Group Design Standards* if at least one of those conditions is not satisfied.
- **A CACE estimate from an RCT with high attrition** is rated *Meets WWC Group Design Standards With Reservations* if it satisfies three conditions: no clear violations of the exclusion restriction, sufficient instrument strength, and a baseline equivalence requirement. It is rated *Does Not Meet WWC Group Design Standards* if at least one of those conditions is not satisfied.

The review process for CACE estimates is outlined in figure II.6. The following sections provide details on the procedures for assigning ratings to CACE estimates. Section II.3 describes the method for determining whether an RCT has low or high attrition when rating CACE estimates. Sections II.4 and II.5 then describe the procedures for rating CACE estimates from RCTs with low and high attrition, respectively.

¹⁰ Another assumption required for the internal validity of CACE estimates is called monotonicity (Angrist et al., 1996). Under this assumption, anyone who would take up the intervention if assigned to the comparison condition would also do so if assigned to the intervention condition. In other words, it is assumed that there are no individuals who would take up the intervention if assigned to the comparison condition but would not take up the intervention if assigned to the intervention condition. This assumption is not directly verifiable. However, it seems at least as plausible as other unverifiable assumptions that are needed for ITT impacts to attain causal validity, such as the assumption that each subject's outcome is unaffected by the treatment status of other subjects. Therefore, these standards assume that monotonicity is satisfied.

Figure II.6. Review process for studies that report a complier average causal effect estimate

Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in chapter V, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

3. Calculating attrition when rating CACE estimates

When rating CACE estimates, the basic approach to determining whether attrition is low or high will follow the usual attrition standard for RCTs (see section II.A). In particular, both overall and differential attrition must be calculated. Table II.1 will then determine whether the combination of overall and differential attrition is considered low or high.

However, the specific method for calculating attrition rates when rating CACE estimates is different from the method used when rating ITT estimates. When rating ITT estimates, the overall attrition rate is the fraction of the entire randomly assigned sample that did not contribute outcome data to the final analysis. Likewise, the differential attrition rate is the difference in attrition rates between the entire assigned intervention group and entire assigned comparison group. It is appropriate to measure attrition for the entire sample when rating ITT estimates.

because those estimates are intended to represent how assignment to the intervention would, on average, affect all subjects.

In contrast, a CACE estimate represents the average effect of taking up the intervention for compliers only. Accordingly, when rating a CACE estimate, the WWC will calculate overall and differential attrition rates that pertain specifically to compliers. Because compliers cannot be directly identified, as discussed previously, the attrition rates for compliers likewise cannot be directly calculated. Instead, the attrition rates must be estimated on the basis of specific assumptions, discussed next.

For the usual scenario in which there are two assigned groups—the intervention group, denoted by $Z = 1$, and the comparison group, denoted by $Z = 0$ —the differential attrition rate for compliers Δ^{complier} will be estimated as

$$[\text{II.1}] \quad \hat{\Delta}^{\text{complier}} = \frac{\bar{A}_{1,\text{ran}} - \bar{A}_{0,\text{ran}}}{\bar{D}_{1,\text{ran}} - \bar{D}_{0,\text{ran}}},$$

where $\bar{A}_{z,\text{ran}}$ is the attrition rate in the assigned group $Z = z$, and $\bar{D}_{z,\text{ran}}$ is the fraction of the assigned group $Z = z$ that took up the intervention. The numerator of equation II.1 is the differential attrition rate that the WWC calculates when rating ITT estimates, and the denominator is the difference in take-up rates between assigned groups. Equation II.1 provides a consistent estimate of the differential attrition rate for compliers under the assumption that attrition rates for always-takers and never-takers do not differ by assigned status. More generally, equation II.1 provides a conservative, upper-bound estimate of the differential attrition rate for compliers under the assumption that differential attrition rates for always-takers and never-takers, if nonzero have the same sign as the differential attrition rate for compliers. The WWC regards the latter assumption as reasonable and realistic; it is difficult to identify scenarios in which assignment to an intervention would influence attrition patterns in opposite ways for always-takers and never-takers.¹¹

To calculate the overall attrition rate for compliers, we will calculate the attrition rate for compliers in the intervention and comparison groups separately, and then take a weighted average of the two attrition rates, with weights equal to group size. Let \bar{A}_{zd} be the observed attrition rate for people with assignment status $Z = z$ and take-up status $D = d$, with $D = 1$ denoting receipt of the intervention and $D = 0$ denoting nonreceipt. Following Imbens and

¹¹ In most cases, attrition is due to missing outcome data. Less frequently, attrition may be due to missing data on take-up status. If some members of the randomly assigned sample are missing take-up status, then the WWC will not have all of the information needed for calculating the denominator of equation II.1. In this case, we assume a worst-case scenario, in which individuals in the intervention group with missing take-up status truly did not take up the intervention, and individuals in the comparison group with missing take-up status truly took up the intervention. This worst-case scenario minimizes the denominator in equation II.1 and, therefore, leads to an upper-bound for the differential attrition rate.

Rubin (1997b), the attrition rate for compliers in the comparison group R_0^{complier} will be estimated as¹²

$$[\text{II.2}] \quad \hat{R}_0^{\text{complier}} = \frac{(1-\bar{D}_{0,\text{ran}})\bar{A}_{00} - (1-\bar{D}_{1,\text{ran}})\bar{A}_{10}}{\bar{D}_{1,\text{ran}} - \bar{D}_{0,\text{ran}}}.$$

The attrition rate for compliers in the intervention group $\hat{R}_1^{\text{complier}}$ will then be estimated as

$$[\text{II.3}] \quad \hat{R}_1^{\text{complier}} = \hat{R}_0^{\text{complier}} + \hat{\Delta}^{\text{complier}}.$$

The overall attrition rate $\hat{R}_{\text{overall}}^{\text{complier}}$ will then be calculated as

$$[\text{II.4}] \quad \hat{R}_{\text{overall}}^{\text{complier}} = \frac{\hat{R}_1^{\text{complier}} N_1 + \hat{R}_0^{\text{complier}} N_0}{N_1 + N_0},$$

where N_1 and N_0 are the number of sample members randomly assigned to the intervention and comparison groups, respectively.

The procedure described thus far in this section is equivalent to using the units of analysis to estimate a 2SLS regression in which attrition—specifically, a binary variable indicating whether a subject was included in the final analysis sample—is the outcome, a take-up indicator is the endogenous independent variable, and an indicator for assignment to the intervention group serves as the instrumental variable. The estimated coefficient on the take-up indicator is equivalent to the differential attrition rate shown in equation II.1, and the WWC will use the result from this 2SLS regression as the measure of differential attrition when provided.

If there are three or more groups to which each sample member could be randomly assigned, then the procedure we will follow is likewise equivalent to estimating a 2SLS regression in which attrition is the outcome, a take-up indicator is the endogenous independent variable, and a set of assigned group indicators, one for each group except an omitted reference group, constitutes the instrumental variables. In this procedure, we will first order the assigned groups from the lowest to the highest take-up rate. For each comparison between consecutively ordered groups, we will apply equations II.1 through II.4 to obtain differential and overall attrition rates for compliers relevant to that comparison (that is, for subjects who are induced to take up the intervention by being assigned to the higher-ordered group instead of the lower-ordered group). We will then take a weighted average of both the overall and differential attrition rate across those different comparisons, with weights specified in Imbens and Angrist (1994). Appendix C provides formulas for those weights.

¹² The intuition behind equation II.2 is roughly as follows. Members of the assigned comparison group who do not take up the intervention consist of a mix of compliers and never-takers. Starting from the attrition rate for this mixed group, the first term in the numerator of the equation II.2, we remove the contribution coming from comparison-group never-takers, which is assumed to be equivalent to the observed attrition rate of never-takers in the intervention group, the second term in the numerator of the equation II.2. The resulting difference is an estimate of the attrition rate for comparison-group compliers.

4. Procedures for rating CACE estimates when attrition is low

A CACE estimate from a low-attrition RCT is rated *Meets WWC Group Design Standards Without Reservations* if it satisfies two criteria: no clear violations of the exclusion restriction and sufficient instrument strength. If at least one of those criteria is not met, then the CACE estimate is rated *Does Not Meet WWC Group Design Standards*. Next, we describe the two criteria in detail. The conceptual background for these criteria is available in appendix C.

Criterion 1: No clear violations of the exclusion restriction

For a CACE estimate to have no clear violations of the exclusion restriction, a necessary condition is that the study must report a definition of take-up that is the same across assigned groups. Moreover, the WWC's lead methodologist for a review has the discretion to determine that a study fails to satisfy the exclusion restriction as a result of a situation in which assignment to the intervention can materially influence the behavior of subjects even if they do not take up the intervention. For example, the exclusion restriction would be violated if subjects assigned to the intervention group received offers to convince them to enroll in the comparison group instead. See appendix C for additional discussion of violations of the exclusion restriction.

Criterion 2: Sufficient instrument strength

Depending on the number of instruments, a CACE estimate must report a first-stage F statistic—the F statistic for the joint significance of the instruments in the first-stage equation—at least as large as the minimum required level shown in table II.7. The minimum required levels are based on Stock and Yogo's (2005) derivations on the minimum first-stage F statistic needed to ensure that the actual type I error rate is unlikely to exceed 0.10 for a t test whose assumed type I error rate is 0.05.¹³ When there is one instrument, authors may report a t statistic instead. In this case, the F statistic is equal to the square of the t statistic.

When baseline covariates are included in the 2SLS regression, the first-stage F statistic assesses the joint significance of the instruments in the first-stage equation *while controlling for the baseline covariates*. In such cases, the F statistic should only reflect the significance of the instruments, and not the significance of the baseline covariates. If the unit of assignment differs from the unit of analysis, then the study must report first-stage F statistics after adjusting for clustering.

In a limited set of circumstances, the WWC will be able to calculate the first-stage F statistic even if this statistic is not reported by the study and cannot be obtained through an author query. Specifically, in the case of a study with no clustering and one instrumental variable that distinguishes a single intervention group and a single comparison group, the WWC can obtain a conservative, lower-bound value for the first-stage F statistic if information is available on the take-up rate for analysis sample members in the intervention group ($\bar{D}_{1,an}$), the take-up rate for analysis sample members in the comparison group ($\bar{D}_{0,an}$), the number of analysis sample

¹³ Specifically, the minimum required first-stage F statistic is the critical value for rejecting the null hypothesis that the instruments are weak enough to yield type I error rates exceeding 0.10. See Stock and Yogo (2005) for details. Although it is common for researchers to use a rule of thumb that the F statistic must exceed 10, Table II.7 imposes a stronger requirement. Stock and Yogo's (2005) analyses are a refinement and improvement to the Staiger-Stock (1997) rule of thumb, which states that instruments with a first-stage F value less than 10 should be deemed weak.

members in the intervention group ($N_{1,an}$), and the number of analysis sample members in the comparison group ($N_{0,an}$). The first-stage F statistic is represented as

$$[\text{II.5}] \quad \frac{(\bar{D}_{1,an} - \bar{D}_{0,an})^2}{\frac{\bar{D}_{1,an}(1-\bar{D}_{1,an})}{N_{1,an}} + \frac{\bar{D}_{0,an}(1-\bar{D}_{0,an})}{N_{0,an}}},$$

which is a lower-bound value because it does not take into account precision gains from controlling for other covariates in the first-stage equation.

Table II.7. First-stage F statistic thresholds for satisfying the criterion of sufficient instrument strength

Number of instruments	Minimum required first-stage F statistic	Number of instruments	Minimum required first-stage F statistic
1	16.38	16	52.77
2	19.93	17	55.15
3	22.30	18	57.53
4	24.58	19	59.92
5	26.87	20	62.30
6	29.18	21	64.69
7	31.50	22	67.07
8	33.84	23	69.46
9	36.19	24	71.85
10	38.54	25	74.24
11	40.90	26	76.62
12	43.27	27	79.01
13	45.64	28	81.40
14	48.01	29	83.79
15	50.39	30	86.17

Source: Stock and Yogo (2005).

If a CACE estimate does not have an associated first-stage F statistic reported in the study, then the WWC will attempt to obtain it through an author query. If the authors do not provide this statistic after being queried, then the WWC will try to calculate the first-stage F statistic using the formula above, provided that there is only one instrumental variable and no clustering. If none of these options enables the first-stage F statistic to be identified, then the study does not demonstrate sufficient instrument strength and is rated *Does Not Meet WWC Group Design Standards*.

5. Procedures for rating CACE estimates when attrition is high

A CACE estimate from a high-attrition RCT is rated *Meets WWC Group Design Standards With Reservations* if it satisfies three criteria: no clear violations of the exclusion restriction, sufficient instrument strength, and a baseline equivalence requirement. If at least one of those criteria is not satisfied, then the CACE estimate is rated *Does Not Meet WWC Group Design Standards*.

The first two criteria are identical to those discussed in section II.C.4 for RCTs with low attrition. The remainder of this section describes the third criterion, the baseline equivalence requirement.

The baseline equivalence requirement for CACE estimates in high-attrition RCTs follows the basic elements of the baseline equivalence requirement described in step 3 of section II.A. For each baseline characteristic specified in the review protocol, we will calculate a difference between intervention and comparison group members in the analytic sample. If the reported difference is greater than 0.25 standard deviation in absolute value, then the baseline equivalence is not satisfied. If the difference is between 0.05 standard deviation and 0.25 standard deviation, then the analysis must control for the baseline characteristic in the 2SLS regression. Differences of less than or equal to 0.05 require no statistical adjustment (see table II.2).

However, the specific method for calculating a baseline difference when rating CACE estimates is different from the usual method used when rating QEDs or ITT estimates from high-attrition RCTs. The usual method assesses the degree of imbalance between groups in the entire analytic sample. However, for the purpose of rating CACE estimates, it is necessary to assess the degree of imbalance between groups *only among compliers* in the analytic sample.

For each characteristic X specified in the review protocol, we will use the following approach to calculate the baseline difference between compliers in the intervention and comparison groups within the analytic sample. Let $\bar{X}_{z,an}$ be the mean of the characteristic for members of the analytic sample with assigned status $Z = z$, and let $\bar{D}_{z,an}$ be the take-up rate among analytic sample members with assigned status $Z = z$. We will estimate the baseline difference among compliers as

$$[II.6] \quad \hat{\beta}^{complier} = (\bar{X}_{1,an} - \bar{X}_{0,an}) / (\bar{D}_{1,an} - \bar{D}_{0,an}),$$

and then express this difference in standard deviation units, with standard deviations calculated in the usual way, based on the pooled analytic sample.

The numerator of equation II.6 is the baseline difference that the WWC calculates when rating ITT estimates from high-attrition RCTs, and the denominator is the difference in take-up rates between the intervention and comparison groups in the analytic sample. This equation is justified by the same type of assumption that underlies the differential attrition rate calculation in equation II.1. Specifically, equation II.6 provides a conservative, upper-bound estimate of the baseline difference for compliers under the assumption that baseline differences for always-takers and never-takers, if nonzero, have the same sign as the baseline difference for compliers.

In fact, because attrition is the key source of bias that can lead to baseline differences in RCTs, assumptions about attrition behavior (from chapter II) shape what types of assumptions about baseline differences are reasonable. Baseline differences emerge when intervention group members who leave the study are different from comparison group members who leave the study, resulting in a baseline imbalance between groups among those who remain in the study. Stated differently, baseline differences emerge when assignment to the intervention is associated with the composition of people who stay or leave. The approach to calculating attrition, explained in section II.C, was built on the notion that assignment to the intervention is unlikely to have opposite effects on attrition rates for different subpopulations. By similar logic,

assignment to the intervention is unlikely to have opposite effects on the *types* of sample members who leave the study in different subpopulations. For this reason, the WWC finds it reasonable and realistic to assume that baseline differences have the same sign for always-takers, compliers, and never-takers, justifying the use of equation II.6.

If there are three or more groups to which each sample member could be randomly assigned, then we will first order the assigned groups from lowest to highest take-up rate, calculate baseline differences in the analytic sample between compliers of consecutively ordered groups, and take a weighted average of those baseline differences (Imbens & Angrist, 1994). See appendix C for details.

III. Regression discontinuity designs

Researchers use RDDs when education-related interventions are made available to individuals or groups on the basis of how they compare with a cutoff value on some known measure. Students may be assigned, for example, to a summer school program if they score below a cutoff value on a standardized test, or schools may be awarded a grant based on their score on an application. The variable used to assign subjects to the intervention is commonly referred to as the “forcing,” “assignment,” or “running” variable.

The effects provide consistent estimates of the local average impacts and are comparable with traditional group design trials. Under typical RDD methodology, the effect of an intervention is estimated as the difference in mean outcomes between intervention and comparison group members at the cutoff, adjusting statistically for the relationship between the outcomes and the variable used to assign subjects to the intervention. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and the difference in these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. Stated differently, an effect is said to have occurred if there is a “discontinuity” in the two regression lines at the cutoff. This estimate pertains to average intervention effects for subjects right at the cutoff. RDDs generate asymptotically unbiased estimates of the effect of an intervention if the relationship between the outcome and forcing variable is modeled appropriately (defined in Standard 4 next) and the forcing variable was not manipulated, either behaviorally or mechanically, to influence assignment to the intervention group.

This chapter presents criteria under which estimates of effects from RDD studies can be rated *Meets WWC RDD Standards Without Reservations* and the conditions under which they can be rated *Meets WWC RDD Standards With Reservations*. These standards apply to both “sharp” and “fuzzy” RDDs, defined in section III.C. We provide standards for studies that report a single RDD impact (section III.C), standards for studies that report multiple impacts (section III.D), and standards for studies that report pooled or aggregate impacts (section III.E). As is the case in RCTs, clusters of students—such as schools, classrooms, or any other group of multiple individuals that have the same value of the assignment variable—might be assigned to intervention and comparison groups, and so we provide standards for cluster-assignment studies (section III.F). While the standards are focused on assessing the causal validity of impact estimates, we also describe two reporting requirements (sections III.G and III.H) focused on reporting accurate standard errors.

A. Assessing whether a study is eligible for review as a regression discontinuity design

A study is eligible for review as an RDD study if it meets the following criteria:

- Treatment assignments are based on a numerical forcing variable; subjects with numbers at or above a cutoff value, or at or below that value, are assigned to the intervention group whereas subjects with scores on the other side of the cutoff are assigned to the comparison group. For example, an evaluation of a tutoring program could be classified as an RDD if students with a reading test score at or below 30 are admitted to the program and students with a reading test score above 30 are not. As another example, a study examining the impacts of grants to improve teacher training in local areas could be

considered an RDD if grants are awarded to only those sites with grant application scores that are at least 70. In some instances, RDDs may use multiple criteria to assign the treatment to subjects. For example, a student may be assigned to an afterschool program if the student's reading score is below 30 or the student's math score is below 40. Studies that use multiple assignment variables or cutoffs with the same sample are eligible for review under these standards only if they use a method described in the literature (for example, in Reardon and Robinson [2012] or Wong, Steiner, and Cook [2013]) to reduce those variables to a single assignment variable or analyze each assignment variable separately. If a study does not do this (for example, if it uses the response surface method described by Reardon and Robinson [2012]), then it is not currently eligible for review under these standards. As with RCTs, noncompliance with treatment assignment is permitted, but the study must still meet the criteria outlined in this chapter to be eligible for a rating of *Meets WWC RDD Standards*.

- The forcing variable is *ordinal*—that is, it has a unique ordering of the values from lowest to highest—and includes a minimum of four or more unique values below the cutoff and four or more unique values above the cutoff. This condition is required to model the relationship between the outcomes and the forcing variable. The forcing variable must never be based on nonordinal categorical variables, such as sex or race. The analyzed data must also include at least four unique values of the forcing variable below the cutoff and four unique values above the cutoff. This is required for eligibility because at least eight data points are required to credibly select bandwidths or functional forms for the relationship between the outcome and the forcing variable.
- The study must not have a *confounding factor* as defined for group design studies in chapter V. A confounding factor is a component of the study design that is perfectly aligned with either the intervention or comparison group. That is, some factor is present for members of only one group and absent for all members in the other group. In particular, the cutoff value of the forcing variable must not be used to assign members of the study sample to interventions other than the one being tested. For example, the income cutoff for determining free or reduced-price lunch status cannot be the basis of an RDD because free or reduced-price lunch data are used as the eligibility criteria for a wide variety of services that also could affect student achievement. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions. A study can examine the combined impact of two or more interventions that all use the same cutoff value; in that case, the study can be eligible for review as an RDD, but the causal statements made must be about the combined impact because the causal effects of each individual intervention cannot be isolated.
- The forcing variable used to calculate impacts must be the *actual* forcing variable, not a proxy or estimated forcing variable. A variable is considered to be a proxy if its correlation with the actual forcing variable is less than 1.

If a study claims to be based on an RDD but does not have these properties, then the study is not eligible for review as an RDD.

B. Possible ratings for studies using regression discontinuity designs

Once a study is determined to be an RDD, the study can receive one of three ratings based on the set of criteria described below and summarized in table III.1.

1. *Meets WWC RDD Standards Without Reservations.* To qualify, a study must completely satisfy each of the five individual standards listed in table III.1.
2. *Meets WWC RDD Standards With Reservations.* To qualify, a study must at least partially satisfy each of the following standards: 1, 4, 5, and either 2 or 3.
3. *Does Not Meet WWC RDD Standards.* A study will receive this rating if it does not at least partially satisfy any of standards 1, 4, or 5, or does not at least partially satisfy both standards 2 and 3.

Table III.1. Regression discontinuity design study ratings

Standard	To be rated <i>Meets WWC RDD Standards Without Reservations</i> , studies must:	To be rated <i>Meets WWC RDD Standards With Reservations</i> , studies must:
1. Integrity of the forcing variable	Completely satisfy this standard.	Partially satisfy this standard.
2. Sample attrition	Completely satisfy this standard.	Partially satisfy at least one of these two standards.
3. Continuity of the relationship between the outcome and the forcing variable	Completely satisfy this standard.	Partially satisfy at least one of these two standards.
4. Functional form and bandwidth	Completely satisfy this standard.	Partially satisfy this standard.
5. Fuzzy RDD	Completely satisfy this standard.	Partially satisfy this standard.

C. Standards for a single regression discontinuity design impact

The standards presented in this section focus on assessing the causal validity of the impact of a single discontinuity in a single ordinal forcing variable on a single outcome. Section III.D describes how to apply these standards in studies with multiple outcomes or samples. Section III.E describes how to apply these standards in studies with multiple impacts on the same outcome.

Standard 1: Integrity of the forcing variable

A key condition for an RDD to produce consistent estimates of effects of an intervention is that there was no systematic manipulation of the forcing variable. This situation is analogous to the nonrandom manipulation of intervention and comparison group assignments under an RCT. In an RDD, manipulation means that scores for some subjects were systematically changed from their true obtained values to influence treatment assignments and the true obtained values are unknown. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to inconsistent impact estimates.

Manipulation is possible if “scorers” have knowledge of the cutoff value and have incentives and an ability to change unit-level scores to ensure that some subjects are assigned to a specific research condition. Stated differently, manipulation could occur if the scoring and treatment assignment processes are not independent. It is important to note that manipulation of the forcing variable is *different* from treatment status noncompliance, which occurs if some intervention group members do not receive intervention services or some comparison group members receive embargoed services.

The likelihood of manipulation will depend on the nature of the forcing variable, the intervention, and the study design. For example, manipulation is less likely to occur if the forcing variable is a standardized test score than if it is a student assessment conducted by teachers who also have input into treatment assignment decisions. Manipulation is also unlikely in cases where the researchers determined the cutoff value using an existing forcing variable, for example, a score from a test that was administered prior to the implementation of the study.

In all RDD studies, the *integrity of the forcing variable* should be established institutionally, statistically, and graphically.

- **Criterion A. *The institutional integrity of the forcing variable must be established by an adequate description of the scoring and treatment assignment process.*** This description must indicate the forcing variable used; the cutoff value selected; who selected the cutoff, for example, researchers, school personnel, and curriculum developers; who determined values of the forcing variable, for example, who scored a test; and when the cutoff was selected relative to determining the values of the forcing variable. This description must show that manipulation was unlikely because scorers had little opportunity or little incentive to change “true” obtained scores in order to allow or deny specific subjects access to the intervention. If there is both a clear opportunity to manipulate scores and a clear incentive—for example, in an evaluation of a math curriculum if a placement test is scored by the curriculum developer after the cutoff is known—then the study does not satisfy this standard.
- **Criterion B. *The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) to establish the smoothness of the density of the forcing variable right around the cutoff.*** This is important to establish because there may be incentives for scorers to manipulate scores to make subjects just eligible for the intervention group, in which case, there may be an unusual mass of subjects near the cutoff. The statistical test must fail to reject the null hypothesis of continuity in the density of the forcing variable at the 5 percent significance level.
- **Criterion C. *The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff.*** There must not be strong evidence of a discontinuity at the cutoff that is obviously larger than discontinuities in the density at other points, although some small discontinuities may arise when the forcing variable is discrete.

A study can satisfy or partially satisfy this standard if it meets the relevant criteria in table III.2. A study does not satisfy this standard if fewer than two of the three criteria are satisfied.

Table III.2. Satisfying the integrity of the forcing variable standard (standard 1)

Criterion	To completely satisfy the standard, the RDD study:	To partially satisfy the standard, the RDD study:
A. The institutional integrity of the forcing variable must be established by an adequate description of the scoring and treatment assignment process.	Must satisfy this criterion.	
B. The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	Must satisfy any two of the three criteria (A, B, or C).
C. The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	

Standard 2: Sample attrition

An RDD study must have acceptable levels of overall and differential attrition rates (see section II.A). The samples used to calculate attrition must include all subjects who were eligible to be assigned to the intervention or comparison group using the forcing variable, and not only a subset of those subjects known to the researcher. For example, when age is used to assign students to a prekindergarten program, the assignment mechanism typically applies to all students in a defined geographical region, such as a state or district, and at a specified time, such as when a law was passed, or in the fall of a certain school year. An RDD study that examines the impact of the prekindergarten program using age as the assignment variable could only have acceptable levels of attrition if it can identify the full set of students who were present in the geographical region at the specified time. A study calculating attrition only within an administrative dataset on students enrolled in the state's schools several years after assignment would not meet this requirement because the intervention could have affected whether students remained in the state. Put another way, attrition cannot be assessed unless all subjects who were eligible to be assigned to conditions are known and for all of these subjects, their assigned condition must be known.

However, attrition can be assessed within exogenous subgroups, meaning a subgroup identified using a variable that is exogenous to intervention participation; see the subsection on sample loss that is not considered attrition in step 2 of section II.A. For example, attrition could be assessed separately within each site. Also, attrition can be calculated within a bandwidth

around the cutoff value of the forcing variable. Attrition needs to be assessed separately for each contrast of interest.

The way that attrition rates are calculated determines whether an RDD study satisfies this standard completely or partially. Criterion A lists approaches that must be used for an RDD study to completely satisfy this standard. Criterion B lists other approaches that may be used but only allow an RDD study to partially satisfy this standard. Whereas the approaches in criterion A require the author to either use approved methods for statistically adjusting for the forcing variable or apply an acceptable bandwidth for values of the forcing variable, the approaches in criterion B may not provide as accurate an adjustment for the forcing variable. As a result, the approaches in criterion B could result in measures of overall and differential attrition at the cutoff that are less accurate.

- **Criterion A.** *The reported combination of overall and differential attrition rates must be shown to be low using at least one of the following approaches, which have the potential to adjust for the forcing variable most accurately:*
 - Study authors must report the predicted mean attrition rate at the cutoff estimated using data from below the cutoff and the predicted mean attrition rate at the cutoff estimated using data from above the cutoff. Both numbers must be estimated using a statistical model that controls for the forcing variable using the same approach that was used to estimate the impact on the outcome. Specifically, the impact on attrition must be estimated either (A) using exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (B) using the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. For the purpose of applying this standard, the overall attrition rate will be defined as the average of the predicted mean attrition rates on either side of the cutoff, and the differential attrition rate will be defined as the difference in the predicted mean attrition rates on either side of the cutoff.
 - Study authors must calculate overall and differential attrition for the sample inside the bandwidth used for the impact analysis, with or without adjusting for the forcing variable. Although authors do not need to adjust for the forcing variable using this approach, other than by applying the bandwidth, the value of the forcing variable must be known for all subjects so that the bandwidth can be applied.
- **Criterion B.** *The reported combination of overall and differential attrition rates must be shown to be low when calculated using one of the following approaches, which may not provide as accurate an adjustment for the forcing variable as one of the two approaches outlined under criterion A.*
 - Study authors can calculate overall and differential attrition for the entire research sample, adjusting for the forcing variable.
 - Study authors can calculate overall and differential attrition for the entire research sample without adjusting for the forcing variable.

If authors calculate overall and differential attrition both ways—that is, both with and without adjusting for the forcing variable—the WWC will review both and assign the highest possible rating to this part of the study design. Note that approaches should not be mixed; that is, if the rating is based on an overall attrition rate calculated without an adjustment for the forcing

variable, then the differential attrition rate should also be unadjusted. Unlike the approaches in Criterion A, it is possible to assess attrition using the full research sample even when the value of the forcing variable is unknown for some subjects, as long as the assigned conditions of all subjects is known.

A study can satisfy or partially satisfy this standard if it meets the relevant criteria in table III.3. A study does not satisfy this standard if attrition information is not available or if neither of the criteria in the table are met.

Table III.3. Satisfying the attrition standard (standard 2)

Criterion	To completely satisfy the standard, the RDD study:	To partially satisfy the standard, the RDD study:
A. The reported combination of overall and differential attrition rates is low using an approach among those that have the potential to most accurately adjust for the forcing variable.	Must satisfy this criterion.	Does not need to satisfy this criterion.
B. The reported combination of overall and differential attrition rates is low when calculated using an approach among those that may not provide as accurate an adjustment for the forcing variable.	Does not need to satisfy this criterion.	Must satisfy this criterion.

Standard 3: Continuity of the relationship between the outcome and the forcing variable

To obtain a consistent impact estimate using an RDD, there must be evidence that in the absence of the intervention, there would be a smooth relationship between the outcome and the forcing variable at the cutoff score. This condition is needed to ensure that any observed discontinuity in the outcomes of intervention and comparison group subjects at the cutoff can be attributed to the intervention.

This smoothness condition cannot be checked directly, although two indirect approaches could be used. The first approach is to test whether, conditional on the forcing variable, key *baseline* covariates that are correlated with the outcome variable (as identified in the review protocol for the purpose of establishing equivalence) are continuous at the cutoff. This means that the intervention must have no impact on baseline covariates at the cutoff. Particularly important baseline covariates for this analysis are preintervention measures of the key outcome variables, such as pretests in the case of achievement outcomes.

The second approach for assessing the smoothness condition is to use statistical tests or graphical analyses to examine whether there are discontinuities in the outcome-forcing variable relationship at values away from the cutoff. This process involves testing for impacts at values of the forcing variable where there should be no impacts, such as the medians of points above or below the cutoff value (Imbens & Lemieux, 2008). The presence of such discontinuities would imply that the relationship between the outcome and the forcing variable at the cutoff may not be truly continuous, suggesting that observed impacts at the cutoff may not be due to the intervention.

Three criteria determine whether a study satisfies this standard.

- **Criterion A. Baseline equivalence on key covariates, as identified in the review protocol, must be established at the cutoff value of the forcing variable.** This involves calculating an impact at the cutoff on the covariate of interest, and the study must either (1) use exactly the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (2) use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. Authors may exclude sample members from this analysis for reasons that are clearly exogenous to intervention participation. For example, authors may calculate baseline equivalence using only data within the bandwidth that was used to estimate the impact on the outcome. The burden of proof falls on the authors to demonstrate that any sample exclusions were made for exogenous reasons.

The baseline equivalence standards for group designs apply to the results from this analysis; see chapter II of this handbook. Specifically, if the impact for any covariate is greater than 0.25 standard deviation in absolute value, based on the variation of that characteristic in the pooled sample, this criterion is not satisfied. If the impact for a covariate is between 0.05 standard deviation and 0.25 standard deviation, the statistical model used to estimate the average treatment effect on the outcome must include a statistical adjustment for that covariate to satisfy this criterion. Differences of less than or equal to 0.05 require no statistical adjustment.

For dichotomous covariates, authors must provide the predicted mean covariate value—that is, the predicted probability—at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff. Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the covariate at the cutoff. These predicted probabilities are needed so that WWC reviewers can transform the impact estimate into standard deviation units.

If the attrition standard is at least partially satisfied, then the equivalence criterion can be demonstrated using data not in the analytic sample, such as data from a different year, cohort, or site. However, all other requirements specified above apply, including using an acceptable bandwidth and/or functional form, and excluding sample members only for clearly exogenous reasons. The review leadership team, in consultation with content experts, has discretion to determine that the sample is too different from the context in the study sample to satisfy this criterion.

If the attrition standard is not met, this analysis must be conducted using only subjects with nonmissing values of the key outcome variable used in the study. Exogenous exclusions from that sample are allowed. For example, subjects outside of an acceptable bandwidth can be excluded.

- **Criterion B. There must be no evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.** An example of a “satisfactory explanation” is that the discontinuity corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value. Another example could be a known structural property of

the assignment variable, for example, if the assignment variable is a construct involving the aggregation of both continuous and discrete components. The graphical analysis—such as a scatter plot of the outcome and forcing variable using either the raw data or averaged/aggregated data within bins/intervals—must not show a discontinuity at any forcing variable value within the bandwidth (or, for the full sample if no bandwidth is used) that is larger than two times the standard error of the impact estimated at the cutoff value, unless a satisfactory explanation of that discontinuity is provided. (The standard error at the cutoff value is used because authors may not report the standard error at the point of the observed discontinuity.)

- **Criterion C.** *There must be no evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.* The statistical tests must use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome and be conducted for at least four values of the forcing variable below the cutoff and four values above the cutoff; these values can be either within or outside the bandwidth. At least 95 percent of the estimated impacts on the outcome at other values of the forcing variable must be statistically insignificant at the 5 percent significance level. For example, if impacts are estimated for 20 values of the forcing variable, then at least 19 of them must be statistically insignificant.¹⁴

A study can satisfy or partially satisfy this standard if it meets the relevant criteria in table III.4. A study does not satisfy this standard if criterion A is not satisfied, or if both criteria B and C are not satisfied.

Table III.4. Satisfying the continuity of the relationship between the outcome and the forcing variable standard (standard 3)

Criterion	To completely satisfy the standard, the RDD study:	To partially satisfy the standard, the RDD study:
A. Baseline equivalence on key covariates	Must satisfy this criterion.	Must satisfy this criterion.
B. No evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value	Must satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C. No evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value	Must satisfy this criterion.	

¹⁴ If impacts are estimated for fewer than 20 values of the forcing variable, all of them must be statistically insignificant at the 5 percent significance level.

Standard 4: Functional form and bandwidth

Unlike with RCTs, statistical modeling plays a central role in estimating impacts in an RDD study. The most critical aspects of the statistical modeling are the functional form specification of the relationship between the outcome variable and the forcing variable and the appropriate range of forcing variable values used to select the analysis sample, that is, the *bandwidth* around the cutoff value. Six criteria determine whether a study satisfies this standard.

- **Criterion A.** *The local average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable.* For both bias and variance considerations, it is never acceptable to estimate an impact by comparing the mean outcomes of intervention and comparison group members without adjusting for the forcing variable (even if there is a weak relationship between the outcome and forcing variable).
- **Criterion B.** *The study should use a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.* For example, a bandwidth selection procedure described in an article published in a peer-reviewed journal that describes the procedure and demonstrates its effectiveness would be a justified bandwidth. An article published in an applied journal where the procedure happens to be used does not count as justification. A study that does not use a justified bandwidth does not completely satisfy this standard but could partially satisfy this standard if criterion C is satisfied.
- **Criterion C.** *If the study does not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a “best fit” regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.* For an impact estimate to meet this criterion, the functional form of the relationship between the outcome and forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R -squared.
- **Criterion D.** *The study needs to provide evidence that the findings are robust to varying bandwidth or functional form choices.* At least one of five types of evidence is sufficient to meet this criterion¹⁵:
 - In the case that criterion B applies, the sign and significance of impact estimates must be the same for a total of at least two different justified bandwidths. For example, this criterion would be satisfied if the sign and significance of an impact are the same using a bandwidth selected by cross-validation¹⁶ and a bandwidth selected by the method described in Imbens and Kalyanaraman (2012). Two impact estimates are considered to have the same significance if they are both statistically significant at the

¹⁵ If a study presents more than one type of evidence, and one type shows findings are robust while another type does not, then this criterion is still satisfied. That is, studies are not penalized for conducting more sensitivity analyses.

¹⁶ An implementation of cross-validation for RDD analysis is described by Imbens and Lemieux (2008).

5 percent significance level, or if neither of them is statistically significant at the 5 percent significance level. Two impact estimates are considered to have the same sign if they are both positive, both negative, or if one is positive and one is negative, but neither are statistically significant at the 5 percent significance level.

- In the case that criterion B applies, the sign and significance of impact estimates must be the same for at least one justified bandwidth and at least two additional bandwidths that are not justified.
- In the case that criterion C applies, the sign and significance of impact estimates must be the same using a total of at least two different goodness-of-fit measures to select functional form. For example, this criterion would be satisfied if the impact corresponding to the functional form selected using the Akaike Information Criterion is the same sign and significance as an impact corresponding to the functional form selected using the regression R -squared. Note that both measures may select the same functional form.
- In the case that criterion C applies, the sign and significance of impact estimates must be the same for at least three different functional forms, including the “best fit” regression.
- If the study meets both criteria B and C, then the sign and significance of impact estimates must be the same for the impact estimated within a justified bandwidth and the impact estimated using a “best fit” regression.
- **Criterion E. *The report must include a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.*** The display cannot be obviously inconsistent with the choice of bandwidth and the functional form specification for the analysis. Specifically, if the study uses a particular functional form for the outcome-forcing variable relationship, then the study must show graphically that this functional form fits the scatter plot reasonably well, and if the study uses a local linear regression, then the scatter plot must show that the outcome-forcing variable relationship is indeed reasonably linear within the chosen bandwidth.
- **Criterion F. *The relationship between the forcing variable and the outcome must not be constrained to be the same on both sides of the cutoff.***

A study can satisfy or partially satisfy this standard if it meets the relevant criteria in table III.5. A study does not satisfy this standard if either criterion A or criterion E is not satisfied or if both criteria B and C are not satisfied.

Table III.5. Satisfying the functional form and bandwidth standard (standard 4)

Criterion	To completely satisfy the standard, the RDD study:	To partially satisfy the standard, the RDD study:
A. The local average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable.	Must satisfy this criterion.	Must satisfy this criterion.
B. The study should use a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.	Must satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C. If the study does not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a “best fit” regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.	Does not need to satisfy this criterion.	
D. The study needs to provide evidence that the findings are robust to varying bandwidth or functional form choices.	Must satisfy this criterion.	Does not need to satisfy this criterion.
E. The report must include a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.	Must satisfy this criterion.	Must satisfy this criterion.
F. The relationship between the forcing variable and the outcome must not be constrained to be the same on both sides of the cutoff.	Must satisfy this criterion.	Does not need to satisfy this criterion.

Standard 5: Fuzzy regression discontinuity design

In a *sharp* RDD, all intervention group members receive intervention services and no comparison group members receive services. In a fuzzy regression discontinuity design (FRDD), some intervention group members do not receive intervention services or some comparison group members do receive intervention services, but there is still a substantial discontinuity in the probability of receiving services at the cutoff. In an FRDD analysis, the impact of service receipt is calculated as a ratio. The numerator of the ratio is the RDD impact on an outcome of interest. The denominator is the RDD impact on the probability of receiving services. This analysis is typically conducted using either 2SLS or a Wald estimator. FRDD analysis is analogous to a CACE or local average treatment effect analysis—consequently many aspects of this standard are analogous to the WWC standards for CACE analysis in the context of RCTs.

The internal validity of an FRDD estimate depends primarily on three conditions. The first condition, known as the exclusion restriction, requires that the only channel through which assignment to the intervention or comparison groups can influence outcomes is by affecting take-

up of the intervention being studied (Angrist et al., 1996). When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may be attributable to other factors differing between the intervention and comparison groups. The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels. However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.

The second condition for the internal validity of an FRDD estimate is that the discontinuity in the probability of receiving services at the cutoff needs to be large enough to limit the influence of finite sample bias. The FRDD scenario can be interpreted as an IV model in which falling above or below the cutoff is an instrument for receiving intervention services (the participation indicator). IV estimators will be subject to finite sample bias if there is not a substantial difference in service receipt on either side of the cutoff, that is, if the instrument is “weak” (Stock & Yogo, 2005). FRDD impacts need not be estimated using 2SLS methods—for example, they can be estimated using Wald estimators—but authors must run the first-stage regression of the participation indicator on the forcing variable and the indicator for being above or below the cutoff, and provide either the F statistic or the t statistic from this regression.

The third condition for the internal validity of an FRDD estimate is that two relationships need to be modeled appropriately: the relationship between the forcing variable and the outcome of interest (standard 4) and the relationship between the forcing variable and receipt of services. Ideally, the FRDD impact would be estimated using a justified bandwidth and functional form, where justification is focused on the overall FRDD impact, not just the numerator or denominator separately. Several methods have been discussed in the literature for selecting a justified bandwidth that targets the ratio (such as Calonico, Cattaneo, & Titiunik, 2014; Imbens & Kalyanaraman, 2012). However, in practice authors often use the bandwidth for the numerator of the FRDD, which is consistent with advice from Imbens and Kalyanaraman (2012).¹⁷

Eight criteria determine whether a study satisfies this standard. All eight criteria are waived for impact estimates calculated using a reduced form model (in which the outcome is modeled as a function of the forcing variable, an indicator for being above or below the cutoff, and possibly other covariates, but the *participation* indicator is not included in the model). This type of model is analogous to an ITT analysis in the context of RCTs.¹⁸

¹⁷ Imbens and Kalyanaraman (2012, p. 14) wrote, “In practice, this often leads to bandwidth choices similar to those based on the optimal bandwidth for estimation of only the numerator of the RD estimate. One may therefore simply wish to use the basic algorithm ignoring the fact that the regression discontinuity design is fuzzy.”

¹⁸ An important consideration when interpreting and applying these standards is that they are focused on the causal validity of impact estimates, not on appropriate interpretation of impact estimates. While the reduced form impact estimate may be a valid estimate of the effect of being below (or above) the RDD cutoff, interpreting that impact can be challenging in some contexts. In particular, while the reduced form RDD impact is methodologically analogous to the ITT impact from an RCT, the substantive interpretation can be entirely different. Addressing these interpretive issues is beyond the scope of these standards, but we urge users of these standards to think carefully about interpretation.

- **Criterion A. The participation indicator must be a binary indicator for taking up at least a portion of the intervention.** For example, the participation indicator could be a binary indicator for receiving any positive dosage of the intervention.
- **Criterion B. The estimation model must have exactly one participation indicator.**
- **Criterion C. The indicator for being above or below the cutoff must be a binary indicator for the intervention and comparison groups to which subjects are assigned.**
- **Criterion D. The same covariates, one of which must be the forcing variable, must be included in the analysis that estimates the impact on participation and the analysis that estimates the impact on outcomes.** In the case of 2SLS estimation, this means that the same covariates must be used in the first and second stages.
- **Criterion E. The FRDD estimate must have no clear violations of the exclusion restriction.** Defining participation inconsistently between the assigned intervention and assigned comparison groups would constitute a clear violation of the exclusion restriction. Therefore, the study must report a definition of take-up that is the same across assigned groups. Another violation of the exclusion restriction is the scenario in which assignment to the intervention group changes the behavior of subjects even if they do not take up the intervention itself. In this case, the treatment assignment might have effects on outcomes through channels other than the take-up rate. There must be no clear evidence that assignment to the intervention influenced the outcomes of subjects through channels other than take-up of the intervention.
- **Criterion F. The study must provide evidence that the forcing variable is a strong predictor of participation in the intervention.** In a regression of program participation on a treatment indicator and other covariates, the coefficient on the treatment indicator must report a minimum F statistic of 16 or a minimum t statistic of 4.¹⁹ For FRDD studies with more than one indicator for being above or below the cutoff, see the WWC Group Design Standards for RCTs that report CACE estimates for the minimum required first-stage F statistic.
- **Criterion G. The study must use a local regression or related nonparametric approach in which FRDD impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature.** Ideally, this method would be justified for the FRDD impact estimate, not just the numerator of the FRDD estimate. However, two other approaches are acceptable. First, it is acceptable to use separate bandwidths for the numerator and denominator, if both are selected using a justified approach, such as the IK algorithm applied separately to the numerator and denominator. Second, it is acceptable to use the

¹⁹ Stock and Yogo (2005). The F statistic must be for the instrument only—not the F statistic for the entire first stage regression. If the unit of assignment does not equal the unit of analysis, then the F statistic or t statistic must account for clustering using an appropriate method (such as boot-strapping, hierarchical linear modeling [HLM], or the method proposed by Lee and Card, 2008). Also, in a working paper, Fier, Lemieux, and Marmer (2016) suggested that in the FRDD context, the minimum first-stage F statistic that ensures asymptotic validity of a 5 percent two-sided test is much higher than would be required in a simple IV setting; specifically, they suggest 135. Until a published paper provides an F statistic cutoff that is appropriate for FRDD studies that use a justified bandwidth, the F statistic of 16 will be used as the interim criterion for assessing instrument strength.

bandwidth selected for the numerator if that bandwidth is smaller than or equal to a justified bandwidth selected for the denominator.

- **Criterion H. If Criterion G is not met, the study can still partially satisfy the standard if the FRDD impact is estimated using a bandwidth that is only justified for the numerator, even if it is larger than a bandwidth justified for the denominator.** This criterion is also satisfied if the denominator is estimated using a “best fit” functional form. That is, the functional form of the relationship between program receipt and the forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R -squared.

A study can satisfy or partially satisfy this standard if it meets the relevant criteria in table III.6. A study does not satisfy this standard if any of criteria A–F are not satisfied, or if both criteria G and H are not satisfied.

Table III.6. Satisfying the fuzzy regression discontinuity design standard (standard 5)

Criterion	To completely satisfy the standard, the RDD study:	To partially satisfy the standard, the RDD study:
A. The participation indicator must be a binary indicator	Must satisfy this criterion.	Must satisfy this criterion.
B. The estimation model must have exactly one participation indicator	Must satisfy this criterion.	Must satisfy this criterion.
C. The indicator for being above or below the cutoff must be a binary indicator for the groups	Must satisfy this criterion.	Must satisfy this criterion.
D. The same covariates must be included in (1) the analysis that estimates the impact on participation and (2) the analysis that estimates the impact on outcomes	Must satisfy this criterion.	Must satisfy this criterion.
E. No clear violations of the exclusion restriction	Must satisfy this criterion.	Must satisfy this criterion.
F. Evidence that the forcing variable is a strong predictor of participation in the intervention	Must satisfy this criterion.	Must satisfy this criterion.
G. Local regression or related nonparametric approach with a justified bandwidth	Must satisfy this criterion.	Does not need to satisfy this criterion.
H. Local regression or related nonparametric approach with a bandwidth that is only justified for the numerator or the denominator is estimated using a best fit functional form	Does not need to satisfy this criterion.	Must satisfy this criterion.

D. Applying standards to studies that report multiple impact estimates

Some RDD studies report multiple separate impacts, for example, impacts for different outcomes or subgroups of interest. Each of the standards described above will be applied to each outcome-subgroup combination, resulting in a separate rating for each combination. The overall rating for the study will be the highest rating attained by any outcome-subgroup combination and will apply to only the combination(s) with that rating. In section III.E, we address the special

case of impacts that are pooled or aggregated across multiple combinations of forcing variables, cutoffs, and samples.

E. Applying standards to studies that involve aggregate or pooled impacts

Some RDD studies may report pooled or aggregate impacts for some combinations of forcing variables, cutoffs, and samples. By “pooled impact,” we mean that data from each combination of forcing variable, cutoff, and sample are standardized and grouped into a single dataset for which a single impact is calculated. By “aggregate impact,” we mean a weighted average of impacts that are calculated separately for every combination of forcing variable, cutoff, and sample.

The overall rating for the study will be the highest rated impact—including pooled and aggregate impacts—presented in the study. Authors may improve the rating of a pooled or aggregate impact by excluding combinations of forcing variables, cutoffs, and samples that do not meet WWC RDD standards for reasons that are clearly exogenous to intervention participation. For example, in a multisite study, a site that fails the institutional check for manipulation could be excluded from the aggregate impact, resulting in a higher rating for the aggregate impact. However, potentially endogenous exclusions—those potentially influenced by the intervention—will not improve the rating of an aggregate impact because standards will be applied as if those exclusions were not made. For example, excluding sites that have a high differential attrition rate from an aggregate impact will not improve the rating of that impact because for the purpose of applying the attrition standard, we will include those sites. The burden of proof falls on the authors to demonstrate that any exclusions from the aggregate impact were made for exogenous reasons.

For each impact that is based on a single forcing variable, cutoff, and sample, the standards can be directly applied as stated in section III.C.

For pooled or aggregate impacts that are based on multiple forcing variables, cutoffs, or samples, additional guidance for applying the standards is provided next.

Standard 1: Integrity of the forcing variable

- **Criterion A.** *If the institutional integrity of the forcing variable is not satisfied for any combination of forcing variable, cutoff, and sample that are included in a pooled or aggregate impact, then this criterion is not satisfied for that pooled or aggregate impact.* However, it is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion. For example, if a pooled or aggregate impact is estimated using data from five sites, and the institutional integrity of the forcing variable is not satisfied in one of those five sites, then the pooled or aggregate impact does not satisfy this criterion. However, a pooled or aggregate impact estimated using data from only the four sites for which the institutional integrity of the forcing variable is satisfied would satisfy this criterion.
- **Criterion B.** *For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.* In the case of a pooled impact, applying an appropriate statistical test to the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

- **Criterion C.** *For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.* In the case of a pooled impact, providing a single figure based on the pooled data can also satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

Standard 2: Attrition

In the case of a pooled impact, the attrition standard described in section III.C can be applied directly if the authors calculate and report overall and differential attrition using the pooled sample. Any sample excluded from calculating the pooled or aggregate impact for reasons of endogeneity—that is, because the sample was potentially influenced by the intervention—cannot be excluded from the attrition calculation.

In the case of an aggregate impact, the WWC attrition standard can be applied to the overall and differential attrition rates calculated as weighted averages of the overall and differential rates calculated for each unique combination of forcing variable, cutoff, and sample that contribute to the aggregate impact. Authors must calculate overall and differential attrition for each of those unique combinations in a way that is consistent with the standard described in section III. C, and the weights used in aggregation must be the same weights used to calculate the weighted impact being reviewed. The attrition standard described in section III.C is then applied to the combination of overall and differential attrition based on the weighted average.

Standard 3: Continuity of the relationship between the outcome and the forcing variable

- **Criterion A.** *In the case of a pooled impact, this criterion can be applied as described in section III.C without modification. In the case of an aggregate impact, baseline equivalence can be established by applying the same aggregation approach to the impacts on baseline covariates as is used to aggregate impacts on outcomes.*
- **Criterion B.** *In the case of a pooled impact, this criterion can be applied as described in section III.C without modification. In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples.* Specifically, there must not be evidence of a discontinuity larger than twice the standard error of the impact at any noncutoff value within the bandwidth of any forcing variable for any sample. This means that a graphical analysis must be presented for every combination of forcing variable, cutoff, and sample. In cases where impacts from disjointed—that is, nonoverlapping—samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion, such an exclusion is considered exogenous.
- **Criterion C.** *In the case of a pooled impact, this criterion can be applied as described in section III.C without modification. In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples.* That is, at least 95 percent of estimated impacts at values of the forcing variables other than the cutoffs, across all samples, must be statistically insignificant. In cases where impacts from disjoint samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion; such an exclusion is considered exogenous.

Standard 4: Functional form and bandwidth

In the case of a pooled impact, this standard can be applied as described in section III.C without modification.

In the case of an aggregate impact, criteria A, B, C, E, and F of this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate because they do not satisfy one of those criteria will be treated as attrition. The aggregate impact will receive the lowest rating from among all of these impacts.

Criterion D can be applied only to the aggregate impact. That is, it is sufficient to demonstrate robustness of the aggregate impact—it is not necessary to show robustness of every impact included in the aggregate, although showing robustness for every individual impact is also acceptable.

Standard 5: Fuzzy regression discontinuity design

In the case of a pooled impact, this standard can be applied as described in section III.C without modification.

In the case of an aggregate impact, this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate will be treated as attrition, with two exceptions—impacts may be excluded if they do not meet criterion E or F. The aggregate impact will receive the lowest rating from among all of these impacts.

F. Cluster-assignment regression discontinuity designs

The WWC considers an RDD study to be a cluster-assignment study when individuals are assigned to conditions in groups and the outcome measure is assessed for individuals within clusters. The same two screening conditions for cluster-assignment group design studies apply as are discussed in section II.B. We provide additional criteria for applying the five RDD standards to cluster-assignment RDDs here. These criteria describe how and when to use cluster- or individual-level data to satisfy each RDD standard.

As with cluster group design studies, cluster RDDs can satisfy WWC standards for effects of an intervention on individuals or on clusters. The WWC initially reviews a cluster RDD study for evidence of an intervention’s effect on individuals. If an effect on individuals cannot be credibly demonstrated, then the WWC reviews the evidence of an intervention’s effect on clusters, where changes in the composition of individuals within the clusters may influence the observed effect. When an RDD study satisfies WWC standards for effects of the intervention on individuals, it may be eligible for the highest rating of *Meets WWC RDD Standards Without Reservations*. However, the observed impact estimate in an RDD study that satisfies WWC standards for effects on clusters but not on individuals potentially represents a combination of the effect of the intervention on individuals and a composition effect due to different types of individuals entering intervention and comparison clusters. Therefore, when an RDD satisfies only those WWC standards for effects on clusters, the study is only eligible to be rated *Meets WWC RDD Standards With Reservations*.

Standards 1, 4, and 5

These standards are assessed in the same way whether the study is being reviewed for evidence of an intervention's effect on individuals or on clusters. Each of these standards is assessed using the criteria described in section III.C, using individual-level or cluster-level data. For example, if schools are assigned to condition and the study estimates the impact of the intervention by examining student standardized test data averaged to the school level, then criteria B and C of standard 1 (Integrity of the forcing variable) could be assessed using school-level data or student-level data (the assessment of criterion A does not rely on study data).

Standard 2: Attrition

The attrition standard can be completely or partially satisfied in the review of a cluster RDD for effects on individuals. If the standard is not satisfied in the review for effects on individuals, then it may be partially satisfied (but not completely satisfied) in the review of the study for effects on clusters.

Review of a cluster RDD for effects on individuals

In the review of a cluster RDD for evidence of effects on individuals, individuals who enter clusters after the results of assignment are known may pose a risk of bias. Therefore, the attrition standard includes an assessment of potential risk of bias from joiners. If the analytic sample includes individuals who joined clusters after random assignment and those individuals pose a risk of bias, then the attrition standard can only be partially satisfied, and the highest rating the study can receive is *Meets WWC RDD Standards With Reservations*.

For a cluster-assignment RDD study to *completely satisfy* the attrition standard in the review for evidence of effects on individuals, the study must meet the following three requirements:

- Limit the risk of bias from individuals who entered clusters after assignment as described in step 2 of chapter II.
- Meet the same requirements for completely satisfying the standard using individual-level data within nonattriting clusters, applying an acceptable reference sample as defined in step 3 of chapter II.
- Meet the requirements for completely satisfying the standard as described in III.C using cluster-level data.

To *partially satisfy* the standard in the review for evidence of effects on individuals, the study must meet the following requirements:

- Limit the risk of bias from individuals who entered clusters after assignment as described in step 2 of chapter II.
- Meet the same requirements for completely or partially satisfying the standard using individual-level data within nonattriting clusters, applying an acceptable reference group as defined in step 3 of chapter II.
- Meet the requirements for completely or partially satisfying the standard as described in section II.C using cluster-level data.

Review of a cluster RDD for effects on clusters

In the review of a cluster RDD for evidence of effects on clusters, the study cannot *completely satisfy* the attrition standard because of the risk that impact estimates may in part reflect compositional changes.

To *partially satisfy* the standard in the review of evidence of effects on clusters, the study must meet the following two requirements:

- Meet the requirements for completely or partially satisfying the standard as described in section III.C using cluster-level data.
- Demonstrate that the analytic sample of individuals used to estimate the impact of the intervention is representative of the clusters as described in step 5 of chapter II. The attrition calculations for this representativeness requirement must be performed using an approach that would completely or partially satisfy the RDD attrition standard described in section III.C.

Standard 3: Continuity

The continuity standard can be completely or partially satisfied in the review of a cluster RDD for effects on individuals. If the standard is not satisfied in the review for effects on individuals, then it may be partially satisfied, but not completely satisfied, in the review of the study for effects on clusters.

Review of a cluster RDD for effects on individuals

For a cluster RDD to *completely satisfy* this standard, the study must meet the requirements for satisfying the continuity standard described in section III.C. If the attrition standard is not satisfied in the review for effects on individuals, then criterion A of the continuity standard must be satisfied using the analytic sample of individuals—those who contribute outcome data to the impact analysis. For studies that analyze outcomes aggregated to the cluster level, the analytic sample of individuals are those who contribute outcome data to the cluster-level averages. These requirements can be met using individual-level or cluster-level data.

For a cluster RDD to *partially satisfy* the standard, the study must meet the requirements for partially satisfying the continuity standard described in section III.C. Again, if the attrition standard is not satisfied in the review for effects on individuals, then criterion A of the continuity standard must be satisfied using the analytic sample of individuals.

Review of a cluster RDD for effects on clusters

In the review of a cluster RDD for evidence of effects on clusters, the study cannot *completely satisfy* the continuity standard because of the risk that impact estimates may in part reflect compositional changes.

To *partially satisfy* the standard in the review of evidence of effects on clusters, the study must meet the following requirements:

- Meet the requirements for completely or partially satisfying the continuity standard as described in section III.C, where criterion A of the standard must be satisfied using the analytic sample of clusters if the attrition standard is not satisfied in the review for effects on clusters.
- Demonstrate that the sample of individuals used to assess criterion A of the continuity standard is representative of the clusters as described in step 7 of chapter II.
- Demonstrate that the samples of individuals used to assess criteria B and C of the continuity standard and the analytic sample used to estimate impacts are representative of the clusters as described in step 5 of chapter II. Frequently, the samples used to assess these criteria will be identical to those used to assess impacts, so this representativeness requirement need only be assessed once.

G. Reporting requirement for studies with clustered sample

As is the case in RCTs, clusters of students or other individuals might be assigned in groups to the intervention and comparison conditions. Clustering affects standard errors but does not lead to biased impact estimates, so if study authors do not appropriately account for the clustering of students, a study can still meet WWC RDD standards if it satisfies the standards described above. However, because the statistical significance of findings is used for the rating of the effectiveness of an intervention, when observations are clustered into groups and the unit of assignment, the cluster, differs from the unit of analysis, the individual, study authors must account for clustering using an appropriate method in order for findings reported by the author to be included in the rating of effectiveness. Appropriate methods including boot-strapping, multilevel linear modeling, or the method proposed by Lee and Card (2008). If the authors do not account for clustering, then the WWC will not rely on the statistical significance of the findings from the study.

H. Reporting requirement for dichotomous outcomes

For dichotomous outcomes, study authors must provide the predicted mean outcome—that is, the predicted probability—at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff. Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the outcome at the cutoff. These predicted probabilities are needed in order for findings reported by the author for those outcomes to be included in the rating of effectiveness.

IV. Single-case design studies

These standards are intended to guide WWC reviewers in identifying and evaluating SCDs. If a study is an eligible SCD, it is reviewed using the study rating criteria to determine whether it receives a rating of *Meets WWC SCD Standards Without Reservations*, *Meets WWC SCD Standards With Reservations*, or *Does Not Meet WWC SCD Standards*.²⁰

Eligible SCDs are identified by the following features:

- An individual **case** is the unit of intervention administration and data analysis. A case may be a single participant or a cluster of participants, such as a classroom or school.
- Within the design, the case can provide its own control for purposes of comparison. For **example, the case's series of outcome variables prior to the intervention is compared with** the series of outcome variables during and after receiving the intervention.
- The outcome variable is measured *repeatedly* within and across *different* conditions or levels of the independent variable. These different conditions are referred to as **phases**, such as the first baseline phase, first intervention phase, second baseline phase, and second intervention phase.

The standards for SCDs apply to a wide range of designs, including ABAB designs, multiple baseline designs, alternating and simultaneous intervention designs, changing criterion designs, and variations of these core designs like multiple probe designs. Even though SCDs can be augmented by including one or more independent comparison cases, in this document, the WWC SCD standards address only the core SCDs and are not applicable to the augmented independent comparison SCDs.

Determining a study rating

If the study appears to be an SCD, the following rules are used to determine whether the study's design *Meets WWC SCD Standards Without Reservations*, *Meets WWC SCD Standards With Reservations*, or *Does Not Meet WWC SCD Standards*. In order to meet standards, the following design criteria must be present, as illustrated in figure IV.1:

Data availability

- SCD studies must provide raw data in graphical or tabular format to permit visual analysis of the data to help the WWC assess whether the study meets WWC standards of internal validity for SCDs.

Independent variable

- The independent variable indicating assignment to the intervention must be systematically manipulated; the researcher will determine when and how the independent variable conditions change.

²⁰ For studies that are rated *Meets WWC SCD Standards With or Without Reservations*, an effect size is calculated if it is possible to do so.

Interassessor agreement

- For each case, the outcome variable must be measured systematically over time by more than one assessor. The design needs to collect interassessor agreement (IAA) in each phase and at least 20 percent of the data points in each baseline and intervention condition, and the IAA must meet minimal thresholds. IAA, commonly called interobserver agreement, must be documented on the basis of a statistical measure of assessor consistency. Although there are more than 20 statistical measures to represent IAA (for example, Berk, 1979; Suen & Ary, 1989), commonly used measures include percentage or proportional agreement and Cohen's kappa coefficient, which adjusts for the expected rate of chance agreement (Hartmann, Barrios, & Wood, 2004). According to Hartmann et al., (2004), minimum acceptable values of IAA are at least 0.80, if measured by percentage agreement, and at least 0.60, if measured by Cohen's kappa. The IAA needs to meet these minimum values for each outcome *across all phases and cases*, but not separately for each case or phase. If study does not meet these minimum values for each outcome *across all phases and cases*, then the study is rated *Does Not Meet WWC SCD Standards*.²¹

Residual treatment effects (if applicable)

- Alternating treatment (AT) designs and designs with an intervening third condition are potentially subject to *residual treatment effects*—responses within phases and conditions that are caused by interventions in previous phases and conditions. When there are three or more interventions in an alternating treatment design, the reviewer must ensure that there are no residual treatment effects. If an intervention is judged to have a reasonable likelihood of residual treatment effects, the study is rated *Does Not Meet WWC SCD Standards*.
 - When a review team identifies an eligible alternating treatment design experiment that uses three or more interventions, the review team should ask the content expert to determine whether residual treatment effects are likely given the specific interventions and outcomes in the experiment (the review team can rely on previous approval of similar conditions and outcomes from the content expert; the plausibility of residual effects is not uniquely informed by the data in a given study). The review team should then assign the study for review and pass along the content expert determination to the reviewers. Reviewers should raise any additional concerns they have about residual treatment effects as part of their reviews.
 - In most cases, the plausibility of residual treatment effects is based on theoretical and contextual considerations. Concerns about residual treatments should focus on study design and intervention characteristics, rather than on observed data.
 - If the content expert and reviewer both agree that there are likely to be residual treatment effects, then the study is rated *Does Not Meet WWC SCD Standards* because the measures of effectiveness cannot be attributed solely to the intervention.

²¹ Author queries should be conducted if the authors do not report the total percentage of sessions checked for IAA, whether IAA was checked at least once in each phase for each participant, or the IAA statistic—for example, percentage agreement—was used to demonstrate reliability. An author query also should be conducted if the authors do not specify that IAA data were collected during *each phase and for each case* for an outcome.

If the content expert and reviewer disagree, then review team leadership should revisit the issue with the content expert. If the content expert and reviewer both agree that residual treatment effects are unlikely, then the reviewer should complete the review assuming there are no residual treatment effects.

- Reversal-withdrawal designs, multiple baseline, and multiple-probe designs generally have longer phases than alternating treatment designs, which means more time will pass between the noncontiguous phases that will be compared (for example, between the first B and second A in an ABCAB reversal-withdrawal design); this feature may make residual effects less important even if they are present. If the reviewer and content expert agree that residual effects are unlikely, or are unlikely to be meaningful, then the reviewer(s) should work with the review team leadership and content experts to identify how best to proceed with the review, focusing only on the intervention of interest and the relevant comparison condition when assigning a study rating (that is, ignoring any third or fourth interventions). The alternating treatment design guidance can be used as a foundation.

Other concerns

- **Confounding factor.** The study must not have a *confounding factor* as defined for SCD studies in chapter V.
- **Training phases, if present, cannot overlap.** Once reviewers have determined that the timing of sessions is presented consistently, they should assess concurrence and effects. In order to have concurrence, the cases still in the baseline phase must continue baseline measurement at or after the time point when a preceding case has the first intervention probe after completing their training. In other words, there can be no overlap in the training phases among the cases in the experiment.
 - If this requirement is not met, then there is no concurrence—the design cannot exclude threats to internal validity and should be rated *Does Not Meet WWC SCD Standards* because there are insufficient data to evaluate the attempts to demonstrate an intervention effect.
 - If this requirement is met, the experiment can be rated *Meets WWC SCD Standards*. In addition, when evaluating concurrence in multiple-probe designs, the WWC also requires that “Each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches the prespecified intervention criterion.” When impacts are expected only after complete delivery of the training, the “first receives the intervention” language should be interpreted as the time point when a case has the first intervention probe after completing their training. (Note that some review protocols allow studies to be rated *Meets WWC SCD Standards With Reservations*, even if they do not meet this multiple probe standard.)

Attempts to demonstrate effect over time and data points per phase

- The study must include at least three attempts to demonstrate an intervention effect at three different points in time.²² The three demonstrations criterion is based on professional convention (Horner, Swaminathan, Sugai, & Smolkowski, 2012).
- Depending on the design type, phases must meet criteria involving the number of data points.²³ Failure to meet any of these criteria results in a study rating of *Does Not Meet WWC SCD Standards*.
 - **Reversal or withdrawal (AB).** Must have a minimum of four phases per case with at least five data points per phase to be rated *Meets WWC SCD Standards Without Reservations*. Must have a minimum of four phases per case with at least three data points per phase to be rated *Meets WWC SCD Standards With Reservations*. Any phases based on fewer than three data points will result in the rating of *Does Not Meet WWC SCD Standards* unless otherwise determined by area team leadership.
 - **Multiple baseline and multiple probe.** Must have a minimum of six phases with at least five data points per phase to be rated *Meets WWC SCD Standards Without Reservations*. Must have a minimum of six phases with at least three data points per phase to be rated *Meets WWC SCD Standards With Reservations*. Any phases based on fewer than three data points will result in the rating of *Does Not Meet WWC SCD Standards* unless otherwise determined by area team leadership. The timing of the design's implementation requires a degree of concurrence when the intervention is being introduced. Otherwise, these designs cannot be distinguished from a series of separate AB designs.
 - **Alternating treatment.** Must have a minimum of five data points per baseline or intervention condition and at most two data points per phase to be rated *Meets WWC SCD Standards Without Reservations*. Must have four data points per condition and at most two data points per phase to be rated *Meets WWC SCD Standards With Reservations*. Any phases based on more than two data points will result in the rating of *Does Not Meet WWC SCD Standards* unless otherwise determined by area team leadership. When designs include multiple intervention comparisons—for example, A versus B, A versus C, C versus B—each intervention comparison is rated separately.
 - **Changing criterion.** The reversal or withdrawal (AB) design standards should be applied to changing criterion designs. Each baseline or intervention change or criterion change should be considered a phase change. As such, there should be at least three different criterion changes to establish three attempts to demonstrate an intervention effect. In some studies using this design, the researcher may reverse or change the criterion back to a prior level to further establish that the change in

²² Although atypical, there might be circumstances in which designs without three replications meet the standards. A case must be made by the topic area team leadership based on content expertise, and at least two WWC reviewers must agree with this decision.

²³ If the topic area team leadership determines that there are exceptions to this standard, these will be specified in the topic area or practice guide protocol. For example, extreme self-injurious behavior might warrant a lower threshold of only one or two data points.

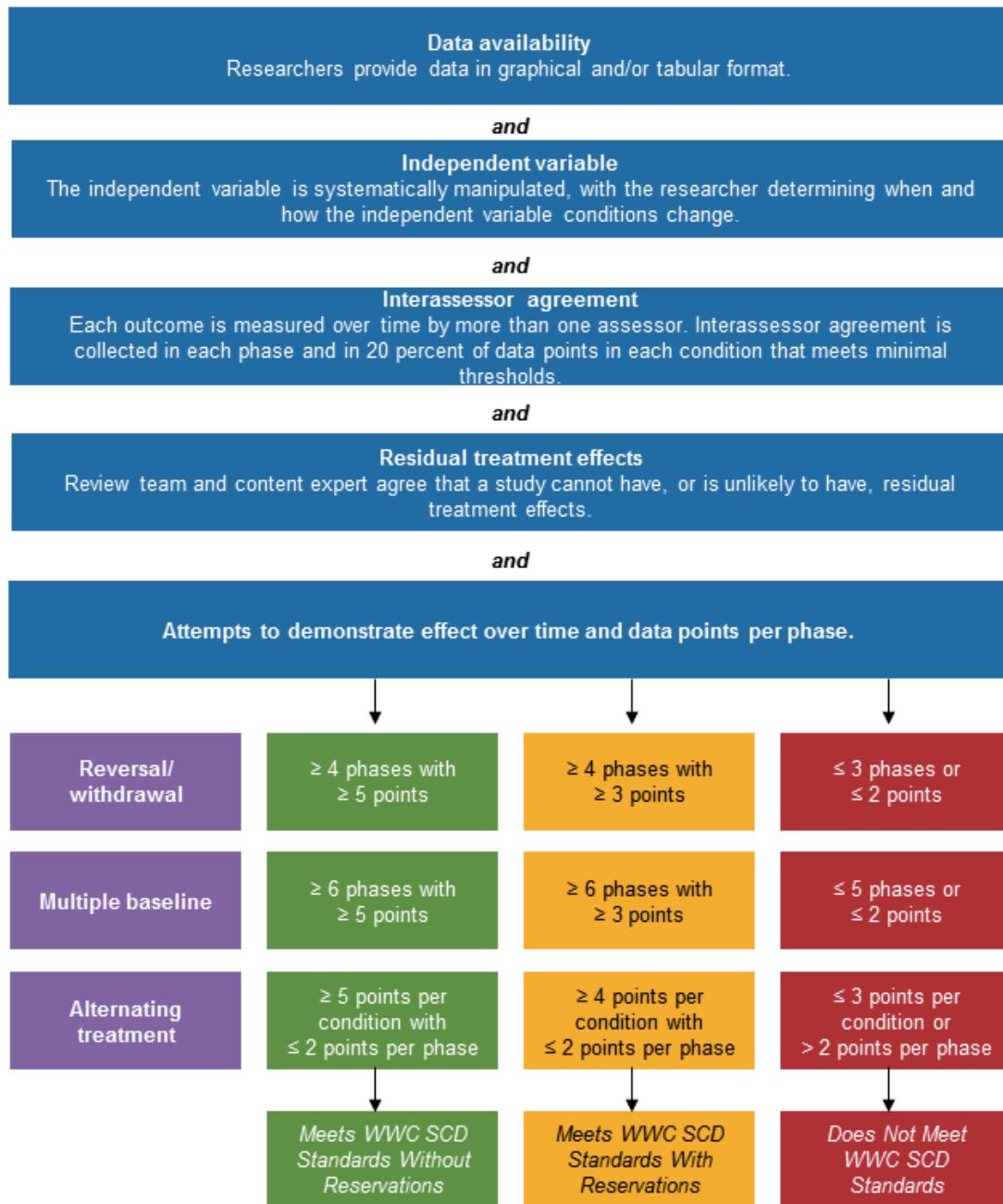
criterion was responsible for the outcomes observed on the dependent variable. This should be considered a phase change, as in the reversal-withdrawal design.

- **Multiple-probe designs.** These designs are a special case of multiple baseline design and must meet additional criteria because baseline data points are intentionally missing.²⁴ Failure to meet any of these results in a study rating of *Does Not Meet WWC SCD Standards*.
 - Initial preintervention data collection sessions must overlap vertically. Within the first three sessions, the design must include three consecutive probe points for each case to be rated *Meets SCD Standards Without Reservations* and at least one probe point for each case to be rated *Meets SCD Standards With Reservations*.
 - Probe points must be available just prior to introducing the independent variable. Within the three sessions just prior to introducing the independent variable, the design must include three consecutive probe points for each case to be rated *Meets WWC SCD Standards Without Reservations* and at least one probe point for each case to be rated *Meets WWC SCD Standards With Reservations*.
 - Each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches the prespecified intervention criterion. This point must be consistent in level and trend with the case's previous baseline points.
- Reversal-withdrawal, multiple-baseline, and multiple-probe designs may have more than the minimum required number of phases required to meet standards, for example, a reversal-withdrawal design with six phases (ABABAB) or a multiple baseline design with four cases where each case has two phases.
 - The reviewer should first conduct the review considering all phases and cases (that is, review the experiment as conducted and reported). If the experiment is rated *Meets WWC SCD Standards With or Without Reservations* when considering all phases and cases, then the reviewer should complete the review without separately considering subsets of phases or cases.
 - If the experiment is rated *Does Not Meet WWC SCD Standards* when considering all relevant phases (for example, because some phases do not have at least three data points), the reviewer should conduct the review considering the subset of consecutive phases (in a reversal-withdrawal design) or consecutive cases (in a multiple baseline or multiple probe design) with enough points and determine whether the subset can meet standards. There may also be multiple rigorous subsets of phases. Reviewers should select the subset aimed at measuring the effectiveness of the intervention of interest. When selecting a subset of phases or cases to review, the ultimate choice should be discussed with review team leadership. Reviewers should document

²⁴ If the topic area team leadership determines that there are exceptions to these standards, then they will be specified in the topic area or practice guide protocol (for example, conditions when stable data patterns necessitate collecting fewer than three consecutive probe points just prior to introducing the intervention or when collecting overlapping initial preintervention points is not possible).

the phases and cases used in the review and the reasons why some may have been excluded from the review. This information will also be documented in WWC products that cite the study.

Figure IV.1. Study rating determinants for single-case designs



V. Nondesign components

In addition to the standards for reviewing eligible studies presented throughout this handbook, two other components may affect a study's rating: outcome requirements and confounding factors.

A. Outcome requirements and reporting

For a finding to be eligible to meet WWC group design standards, it must measure the effect of an intervention on an outcome measure that demonstrates *face validity*, demonstrates *reliability*, is not *overaligned* with the intervention, and (is *collected in the same manner* for both intervention and comparison groups. Standardized tests, in which the same test is given in the same manner to all test takers, are assumed to have face validity and be reliable. Standardized tests have established administration and scoring procedures, often documented in a technical manual. Additionally, behavior outcomes measured using administrative data, such as graduation from high school, school enrollment, and grade retention, are assumed to be reliable because these outcomes are straightforward to measure. Grade point average is also assumed to be reliable if a formula for calculating the measure is specified. Findings based on outcome measures that do not meet all four of these requirements are rated *Does Not Meet WWC Group Design Standards*.

In addition to these four requirements, a study that analyzes imputed outcome data must show that it limits the potential bias from analyzing the imputed outcome data under different assumptions about how the missing data are related to measured or unmeasured factors, as described in section II.C. A study that uses nonresponse weights must also satisfy this requirement. For the study to be eligible to be rated *Meets WWC Group Design Standards With or Without Reservations*, the study must use an acceptable approach to address missing data in the analytic sample; acceptable approaches are listed in section II.C.

Face validity

To show evidence of face validity, a sufficient description of the outcome measure must be provided for the WWC to determine that the measure is clearly defined and the content assessed by the measure aligns with its definition. A measure described as a test of reading comprehension that actually measures reading fluency does not have face validity.

Reliability

The *reliability* requirements aim to set standards for maximum allowable random measurement error, with higher reliability indicating lower measurement error. Internal consistency and test-retest reliability can capture measurement error that results from poor question wording, for example, while inter-rater reliability can capture measurement error that results from coder judgment. Although this random error does not create bias, the error reduces precision and the likelihood of detecting an impact if one actually exists.

Reliability of an outcome measure may be established by meeting the following minimum standards: internal consistency—such as Cronbach's alpha—of .50 or higher; temporal stability and test-retest reliability of 0.40 or higher; or inter-rater reliability—such as percentage

agreement, correlation, or kappa—of .50 or higher. The review protocol may specify higher standards for assessing reliability.

In SCDs, the minimum for percentage agreement—regardless of whether the metric is exact agreement or agreement within 1—is 80 percent (or .80). The minimum kappa or correlation is 0.60. IAA needs to meet these minimum values for each outcome across all phases and cases, but not separately for each case or phase. If the study does not meet these minimum values for each outcome across all phases and cases, then the study is rated *Does Not Meet WWC SCD Standards* because the eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA thresholds.

If study authors do not report that at least 20 percent of the total sessions were checked for IAA and/or that IAA was checked at least once in each phase, then the study is rated *Does Not Meet WWC SCD Standards* because the eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA requirements.

If study authors do not report that IAA data were collected at least once for each phase or case combination, the study is rated *Does Not Meet WWC SCD Standards* because the eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA requirements.

When a study does not report reliability statistics for an outcome measure, the WWC will ask the study authors to provide a statistic. The WWC will also use previously gathered information about reliability of outcomes that are used across studies.

The protocol may also stipulate how to deal with outcome measures that are unlikely to provide reliability information. For example, without quantitatively meeting one of the three reliability standards listed above, an outcome measure may still be deemed reliable if the content expert or lead methodologist for a review determine that responses can be scored by a single coder with low error, such as a multiple-choice test or counts of words spelled correctly. The protocol specifies whether these outcome measures can meet the reliability requirement.

Overalignment

A third requirement of outcome measures is that they not be *overaligned* with the intervention. An outcome measure is overaligned if it contains content or materials provided to subjects in one condition but not the other. When outcome measures are closely aligned with or tailored to the intervention, the study findings may not be an accurate indication of the effect of the intervention. For example, an outcome measure based on an assessment that relied on reading materials or vocabulary words used in the intervention condition but not in the comparison condition likely would be judged to be overaligned.

This rule does not apply when material covered by an outcome measure must be explicitly taught. For example, reciting the alphabet requires being taught the alphabet, but improving reading comprehension does not require focusing on a specific set of reading passages. Put another way, an outcome measure is only overaligned when the content or materials provided to subjects in a single condition might affect scores on the measure through gaming of the outcome measure, familiarity with the format, or other means besides learning educationally relevant material.

The decision about whether a measure is overaligned is made by the review team leadership. In particular, content experts can provide guidance on whether the content assessed in a particular outcome measure is broadly educationally relevant and, thus, not overaligned.

Outcome collection

A fourth requirement of outcome measures is that they be *collected in the same manner* for the intervention and comparison groups. The WWC assumes data were collected in the same manner if no information is provided. However, reviewers look for comments in studies that different modes, timing, or personnel were used for the groups, or measures were constructed differently for the groups. Reviewers may send questions to authors to clarify how data were collected. When outcome data are collected differently for the intervention and comparison groups, study-reported impact estimates will confound differences due to the intervention with those due to differences in the data collection methods. For example, measuring dropout rates based on program records for the intervention group and school administrative records for the comparison group will result in unreliable impact estimates because it will not be possible to disentangle the true impact of the intervention from differences in the dropout rates that are due to the particular measure used. Additionally, when intervention and comparison students are in different districts, grade point average might be calculated differently or be based on different courses in the two groups. If so, it will not be possible to disentangle the impact of the intervention from differences in how the outcome is measured.

B. Confounding factors

In some studies, a component of the study design or the circumstances under which the intervention was implemented are perfectly aligned, or confounded, with either the intervention or comparison group. That is, some factor is present for members of only one group and absent for all members in the other group. Because it is impossible to separate the degree to which an observed effect was due to the intervention and how much was due to the confounding factor, a study with a confounding factor cannot meet WWC standards. In QED studies, confounding is almost always a potential issue due to the selection of a sample, because some unobserved factors may have contributed to the outcome. The WWC accounts for this issue by not allowing a QED study to receive the highest rating.

WWC reviewers must decide whether there is sufficient information to determine that the only difference between the two groups that is not controlled for by design or analysis is the presence of the intervention. If not, there may a confounding factor, and the reviewer must determine whether that factor could affect the outcome separately from the intervention. For the WWC to determine that a confounding factor is present in the study, there must be evidence of its presence. A specific factor that is aligned with the intervention or comparison condition must be identified based on information in the study or obtained from an author query.

This section describes three types of confounding factors: the intervention or comparison group contains a single unit ($n = 1$); the characteristics of the subjects in the intervention or comparison group differ systematically, with no overlap, in ways that are associated with the outcomes; and the intervention is always offered in combination with another intervention and the combined intervention is ineligible for review based on the review protocol and the purpose of the review. Under most review protocols, the WWC will consider studies with the third type of

confounding factor as a reason to screen the study out as ineligible because the study does not examine an intervention with a primary focus aligned with the review protocol, rather than a reason to assign a rating of *Does Not Meet WWC Group Design Standards*. In particular, if a review effort is interested in a specific intervention, and that intervention is combined with another intervention, then the study will not be reviewed for that particular review effort. As it can sometimes be difficult to determine whether something is a confounding factor, the examples that follow describe situations that are and are not confounding factors for each of the three categories.

The intervention or comparison group contains a single unit ($n = 1$).

The most common type of confounding factor occurs when either the intervention or comparison group contains a single study unit—such as a teacher, classroom, or school—and that unit is not present in the other condition. In these situations, there is no way to distinguish between the effect of the intervention and that unit.

- Examples of confounding factors
 - Two schools are randomly assigned, one to each condition.
 - A study has two intervention classrooms and two comparison classrooms, but both intervention classrooms had the same teacher, who had no interaction with the comparison classrooms.
- Examples of similar circumstances that are *not* confounding factors
 - Students are randomly assigned to condition and are all taught by the same teacher in the same school. The WWC does not consider this to be a confounding factor because the same teacher taught both conditions.
 - Schools from three school districts are randomly assigned to a condition. Two of the districts have schools that are represented in both conditions, but all schools in the third district were assigned to a single group. The WWC does not consider this to be a confounding factor because two districts are represented in both groups.
 - A school with unique organization and governance is compared with multiple comparison schools. When the intervention of interest is attending the school, the WWC does not consider this to be a confounding factor because the school and the intervention are the same. However, when the focus of the review is a particular intervention implemented within the school, then the single school would be considered a confounding factor because the effect of the school cannot be distinguished from the effect of the intervention of interest. Additionally, a single school is not a replicable intervention, so if the review protocol states that eligible interventions must be replicable, then the single school would be a confounding factor.

The characteristics of the subjects in the intervention or comparison group differ systematically, with no overlap, in ways that are associated with the outcomes.

Another example of confounding occurs when the characteristics of the subjects in each group differ systematically in ways that are associated with the outcomes. For example, a small group of teachers in a master's program implements the intervention, whereas students in the comparison group are taught by teachers with bachelor's degrees. If the teachers' education is not a component of the intervention—that is, the intervention does not specify that only master's-level teachers can lead the intervention—then it is a potential confounding factor. In

this case, differences in student outcomes between the intervention and comparison groups may be due to the intervention, the higher level of education of the intervention group teachers, or a combination of the two.

When the time period differs for the groups, time is a confounding factor. A design in which groups are defined by cohort is often labeled a *successive-cohort design* or *cohort design*. As an example, an intervention group consists of a cohort of third graders in year Y and the comparison group consists of the previous cohort of third graders in year $Y - 1$. Usually, both cohorts are observed in one school or the same set of schools. In this cohort design, the intervention and comparison conditions are completely aligned with different time periods, and the estimated impact is confounded with any changes that occur between those time periods. These changes—such as new district policies, new personnel, or new state tests—could plausibly affect outcomes. Because many of the changes that occur over time are likely to be unobserved or not reported, the WWC cannot assess how problematic the potential changes are in individual studies.

When there is imperfect overlap in the characteristic between the conditions, this is not a confounding factor. Instead, these situations should be addressed through the usual baseline equivalence requirements specified in the review protocol.

- Examples of confounding factors
 - Intervention students were grade 5 students during the 2014/15 school year, and comparison students were grade 5 students enrolled in the same school during the 2013/14 school year.
 - Intervention students are all in grade 8, and comparison students are all in grade 7.
 - Intervention students are all English learners, and no comparison students are.
- Examples of similar circumstances that are *not* confounding factors
 - Students volunteer to enroll in two different types of mathematics courses: One uses a novel group-based approach as the intervention condition, and one uses a more traditional teacher-directed style as the comparison condition. Some characteristics of students who volunteered for the intervention condition may differ from those who volunteered for the comparison condition—for example, more extroverted students select the group-based program, and more introverted students select the teacher-directed style—but these are not measured by the researcher. The WWC does not consider this to be a confounding factor, but the selection mechanism and potential difference in unmeasured characteristics are reasons that QEDs are limited to a rating of *Meets WWC Group Design Standards With Reservations*, if the baseline equivalence requirement is satisfied on baseline characteristics specified in the review protocol.
 - Classrooms in the intervention condition have much lower rates of students who are eligible for free or reduced-price lunch than classrooms in the comparison condition. The WWC does not consider this to be a confounding factor because there is some overlap in the characteristic between the groups. However, under some review protocols, this difference could be a characteristic on which equivalence must be assessed.

The intervention is always offered in combination with a second intervention and the combined intervention is ineligible for review according to the review protocol and the purpose of the review.

A confounding factor also exists if an intervention is always offered in combination with a second intervention because any subsequent differences in outcomes cannot be attributed solely to either intervention. However, if both interventions are individually eligible for review under the same review protocol, then the WWC may view the combination as a single intervention and report on its effects. Additionally, whereas studies with other types of confounding factors are considered eligible for review but do not meet standards, a study with this type of confounding factor is typically ruled ineligible, and is not assigned a rating, because it does not examine an intervention with a primary focus aligned with the review protocol.

- Example of a confounding factor
 - The focus of the review is a specific software program. Students in the intervention condition were exposed to two software programs, the software program that is the focus of the review and an additional program, but students in the comparison condition were not exposed to either software program.
- Example of a similar circumstance that is *not* a confounding factor
 - The focus of the review is a specific software program. At the same time, everyone in the school is exposed to a second software program, including all intervention and comparison students. The WWC does not consider this to be a confounding factor because all students received the second program, so the only difference between the two groups is the software program of interest.

Confounding factors in SCDs

In SCDs as described in chapter IV, teachers, parents, or peers—collectively labeled *interventionists*—can administer the intervention to study participants. When study participants experience a different interventionist across baseline and intervention phases of the study, the study has a potential confounding factor. This section provides additional guidance for the identification of confounding factors in SCDs.

As it can sometimes be difficult to determine whether something is a confounding factor, the examples below describe situations for which the interventionist is and is not a confounding factor.

- Examples of confounding factors: Participants have a different interventionist across the baseline and intervention phases, noted by underline below.
 - One teacher teaches all cases in the baseline phase, and a different teacher teaches all cases in the intervention phase.

	<i>Baseline</i>	<i>Intervention</i>
Case 1	Teacher 1	<u>Teacher 2</u>
Case 2	Teacher 1	<u>Teacher 2</u>
Case 3	Teacher 1	<u>Teacher 2</u>

- One teacher teaches all cases in the baseline phase, and that same teacher and another teacher (or trainer) teach all cases in the intervention phase.

	<i>Baseline</i>	<i>Intervention</i>
Case 1	Teacher 1	Teacher 1 + <u>Teacher 2</u>
Case 2	Teacher 1	Teacher 1 + <u>Teacher 2</u>
Case 3	Teacher 1	Teacher 1 + <u>Teacher 2</u>

- Examples of similar circumstances that are not confounding factors
 - One teacher teaches all cases in both phases.

	<i>Baseline</i>	<i>Intervention</i>
Case 1	Teacher 1	Teacher 1
Case 2	Teacher 1	Teacher 1
Case 3	Teacher 1	Teacher 1

- Multiple teachers teach different cases; teachers do or do not teach different phases.

	<i>Baseline</i>	<i>Intervention</i>	<i>Baseline</i>	<i>Intervention</i>
Case 1	Teacher 1	Teacher 3	Teacher 1	Teacher 1
Case 2	Teacher 2	Teacher 4	OR	Teacher 2
Case 3	Teacher 2	Teacher 4	Teacher 3	Teacher 3

If a confounding factor is identified, then the study is rated *Does Not Meet WWC SCD Standards* because measures of effectiveness cannot be attributed solely to the intervention.

Appendix A. Assessing bias from imputed outcome data

A. Assessing the bias when the baseline measure is observed for all subjects in the analytic sample

The imputation methods the What Works Clearinghouse (WWC) considers acceptable require assuming that data are missing at random (MAR), which means the missing data depend on measured factors but not on unmeasured factors. If that assumption does not hold, then the impact estimates may be biased. Therefore, quasi-experimental designs (QEDs) and high-attrition randomized controlled trials (RCTs) that use acceptable approaches to impute outcome data must demonstrate that they limit the potential bias from using imputed data to measure impacts. Specifically, potential bias due to deviations from the MAR assumption must not exceed 0.05 standard deviation.

The WWC uses a proxy pattern-mixture modeling approach to estimate the largest possible bias in an impact estimate under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

To bound the bias, we begin by specifying that the probability that we observe an outcome for a given subject is related to the baseline measure and the outcome, which is unmeasured for some cases. This probability in the intervention group ($j = i$) or comparison group ($j = c$) is given by the following function m :

$$[A.1] \quad P_j(x, y) = m\left(\frac{x}{s_x} + \lambda_j \frac{y}{s_y}\right),$$

where x is the baseline measure for a subject, y is the outcome measure for the subject, s_x and s_y are the standard deviations of the baseline and outcome measures, and λ_j measures the deviations from the MAR assumption for group j . When $\lambda_j = 0$, the MAR assumption holds for group j because the missing data depend only on measured baseline data. As λ_j increases, the missingness depends more strongly on the outcome, which may be unmeasured.

Following Andridge and Little (2011), we can write the unmeasured full-sample outcome mean in a group (\bar{y}_j) as a function of the complete case outcome mean (\bar{y}_{jR}), the full-sample and complete case baseline means (\bar{x}_j and \bar{x}_{jR}), and the correlation between the outcome and the baseline measure ρ :

$$[A.2.0] \quad \bar{y}_j = \bar{y}_{jR} + f_j(\rho) \frac{s_y}{s_x} [\bar{x}_j - \bar{x}_{jR}],$$

where the function of ρ is assumed to be:

$$[A.2.1] \quad f_j(\rho) = \frac{\lambda_j + \rho}{\lambda_j \rho + 1}.$$

In many cases, the value of \bar{y}_j will deviate more from the observed mean of \bar{y}_{jR} when there is a larger absolute difference between the full-sample and complete case baseline means. Intuitively, this is because a larger difference means that the subjects with missing outcome data appear different from those with observed outcomes.

When MAR holds, $f_i(\rho) = f_c(\rho) = \rho$ (because $\lambda_i = \lambda_c = 0$), and the expected value of \bar{y}_j is equal to what a researcher would obtain for the full-sample outcome mean when imputing missing values of the outcome measure with predicted values from a regression of the outcome on the baseline measure. But as λ_i or λ_c become larger, the value of $f_j(\rho)$ becomes larger (approaching $1/\rho$), and the outcome mean for the full sample will deviate from the researcher's estimate of the mean using imputed data.

The effect size obtained using an imputation method based on the MAR assumption can be written as the difference in the estimated full-sample intervention and comparison group outcome means with an adjustment for the baseline measure, given by:

$$[A.3.0] \quad g_{MAR} = \frac{1}{s_y} (\{y_{iR} + c[\bar{x}_i - \bar{x}_{iR}]\} - \{y_{cR} + c[\bar{x}_c - \bar{x}_{cR}]\} - c[\bar{x}_i - \bar{x}_c]),$$

where c is the coefficient from a regression of y on x , and is equal to $\rho(s_y/s_x)$.

But this equation can be generalized to the case where the MAR assumption does not hold:

$$[A.3.1] \quad g_{NMAR} = \frac{1}{s_y} \left(\left\{ y_{iR} + f_i(\rho) \frac{s_y}{s_x} [\bar{x}_i - \bar{x}_{iR}] \right\} - \left\{ y_{cR} + f_c(\rho) \frac{s_y}{s_x} [\bar{x}_c - \bar{x}_{cR}] \right\} - c[\bar{x}_i - \bar{x}_c] \right).$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the MAR assumption:

$$[A.4] \quad Bias_y = \frac{1}{s_x} \{ (f_i(\rho) - \rho)[\bar{x}_i - \bar{x}_{iR}] - (f_c(\rho) - \rho)[\bar{x}_c - \bar{x}_{cR}] \}.$$

Because $f_j(\rho)$ is bounded between ρ and $1/\rho$, the largest bias, in absolute value, due to deviations from the MAR assumption is given by the maximum of the values given by the following three equations:

$$[A.5.0] \quad B1 = \omega \left| \frac{\frac{1}{s_x} \frac{1-\rho^2}{\rho}}{s_x} [\bar{x}_c - \bar{x}_{cR}] \right|$$

$$[A.5.1] \quad B2 = \omega \left| \frac{\frac{1}{s_x} \frac{1-\rho^2}{\rho}}{s_x} [\bar{x}_i - \bar{x}_{iR}] \right|$$

$$[A.5.2] \quad B3 = \omega \left| \frac{\frac{1}{s_x} \frac{1-\rho^2}{\rho}}{s_x} [(\bar{x}_i - \bar{x}_{iR}) - (\bar{x}_c - \bar{x}_{cR})] \right|.$$

The bounds in equations A.5.0, A.5.1, and A.5.2 will be calculated using data reported in studies or obtained from authors. The equations include the following data elements described in section II.C: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups (\bar{x}_i , \bar{x}_c , and the standard deviations are used to calculate the pooled within-group standard deviation s_x ²⁵); (b) the means of the baseline measure for the subjects in the analytic sample with observed outcome data,

²⁵ See section IV.A of the *WWC Procedures Handbook, Version 4.1*, for how the WWC calculates the pooled standard deviation.

separately for the intervention and comparison groups (\bar{x}_{iR} , \bar{x}_{cR}); and (c) the correlation between the baseline and the outcome measures (ρ). We have applied a simple correction for bias in the unadjusted Hedges' g effect size when the sample size is small, developed by Hedges (1981), which produces an unbiased effect size estimate by multiplying Hedges' g by a factor of $\omega = [1 - 3/(4N - 9)]$, with N being the total sample size. See appendix E of the *WWC Procedures Handbook, Version 4.1*, for more details.

For simplicity, these bounds were derived for a single baseline measure. If multiple baseline measures were used to form the imputed values in a study, it is acceptable, but not required, to replace the baseline means with the average predicted value of the outcome, that is, the average of the values used to make adjustments to the outcome measure to produce an adjusted mean. In this case, $1/s_x$ is removed from the calculation of the bounds and replaced with $1/s_y$ because the predicted values have units of the dependent variable. Additionally, for outcome domains that require baseline equivalence on multiple baseline measures, it is required that the imputed values adjust for all baseline measures specified in the review protocol and that the bounds are calculated using the average of the predicted values.

B. Assessing the bias when the baseline measure is imputed or missing for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds in equations A.5.0–A.5.2. This is because the means of the baseline measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed outcome data.

Instead, the bounds can be calculated using equations A.10.0–A.10.2. These bounds can be derived by first writing the full sample outcome mean as a weighted sum of the outcome mean for the sample with missing data on the baseline measure, and the sample with observed data on the baseline measure:

$$[A.6.0] \quad \bar{y}_j = \left(\frac{n_j - n_{jx}}{n_j} \right) \bar{y}_{j \sim x} + \left(\frac{n_{jx}}{n_j} \right) \bar{y}_{jx},$$

where n_j is the number of observations in the analytic sample for group j , n_{jx} is the number of observations in the analytic sample for group j with an observed value of the baseline measure, $\bar{y}_{j \sim x}$ is the outcome mean for the observations in the analytic sample for group j missing the baseline measure, and \bar{y}_{jx} is the outcome mean for the remaining members of the analytic sample for group j .

We assume that the analytic sample includes no cases where both the baseline and outcome data are missing, so $\bar{y}_{j \sim x}$ is observed. But \bar{y}_{jx} is not observed because some cases with observed baseline data have missing outcome data. To address this, we write \bar{y}_{jx} as a function of observed measures:

$$[A.6.1] \quad \bar{y}_j = \left(\frac{n_j - n_{jx}}{n_j} \right) \bar{y}_{j \sim x} + \left(\frac{n_{jx}}{n_j} \right) \left(\bar{y}_{jxy} + f_j(\rho) \frac{s_y}{s_x} [\bar{x}_{jx} - \bar{x}_{jxy}] \right),$$

where \bar{y}_{jxy} is the outcome mean for the observations in the complete case analytic sample for group j observed at both baseline and for the collection of outcomes, \bar{x}_{jxy} is the baseline mean for the same sample, and \bar{x}_{jx} is the baseline mean for the sample with observed baseline data but possibly missing outcome data. This equation can be rewritten as:

$$[A.6.2] \quad \bar{y}_j = \bar{y}_{jxy} + \left(\frac{n_j - n_{jx}}{n_j} \right) [\bar{y}_{j \sim x} - \bar{y}_{jxy}] + \left(\frac{n_{jx}}{n_j} \right) f_j(\rho) \frac{s_y}{s_x} [\bar{x}_{jx} - \bar{x}_{jxy}].$$

The effect size obtained using an imputation method based on the MAR assumption ($f_j(\rho) = \rho$) can be written as the difference in the estimated full-sample intervention and comparison group outcome means,²⁶ given by:

$$[A.7] \quad g_{MAR} = \frac{1}{s_y} \left(\left\{ \bar{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i} \right) [\bar{y}_{i \sim x} - \bar{y}_{ixy}] + \left(\frac{n_{ix}}{n_i} \right) c [\bar{x}_{ix} - \bar{x}_{ixy}] \right\} - \left\{ \bar{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c} \right) [\bar{y}_{c \sim x} - \bar{y}_{cxy}] + \left(\frac{n_{cx}}{n_c} \right) c [\bar{x}_{cx} - \bar{x}_{cxy}] \right\} \right).$$

The more general equation that allows deviations from the MAR assumption is given by:

$$[A.8] \quad g_{NMAR} = \frac{1}{s_y} \left(\left\{ \bar{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i} \right) [\bar{y}_{i \sim x} - \bar{y}_{ixy}] + \left(\frac{n_{ix}}{n_i} \right) f_i(\rho) \frac{s_y}{s_x} [\bar{x}_{ix} - \bar{x}_{ixy}] \right\} - \left\{ \bar{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c} \right) [\bar{y}_{c \sim x} - \bar{y}_{cxy}] + \left(\frac{n_{cx}}{n_c} \right) f_c(\rho) \frac{s_y}{s_x} [\bar{x}_{cx} - \bar{x}_{cxy}] \right\} \right).$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the MAR assumption:

$$[A.9] \quad Bias_y = \frac{1}{s_x} \left\{ \left(\frac{n_{ix}}{n_i} \right) (f_i(\rho) - \rho) [\bar{x}_{ix} - \bar{x}_{ixy}] - \left(\frac{n_{cx}}{n_c} \right) (f_c(\rho) - \rho) [\bar{x}_{cx} - \bar{x}_{cxy}] \right\}.$$

The absolute value of this bias is no greater than the maximum of $B1^* - B3^*$:

$$[A.10.0] \quad B1^* = \omega \left| \frac{1}{s_x} \frac{1-\rho^2}{\rho} \left(\frac{n_{ix}}{n_i} \right) [\bar{x}_{ix} - \bar{x}_{ixy}] \right|$$

$$[A.10.1] \quad B2^* = \omega \left| \frac{1}{s_x} \frac{1-\rho^2}{\rho} \left(\frac{n_{cx}}{n_c} \right) [\bar{x}_{cx} - \bar{x}_{cxy}] \right|$$

$$[A.10.2] \quad B3^* = \omega \left| \frac{1}{s_x} \frac{1-\rho^2}{\rho} \left[\left(\frac{n_{ix}}{n_i} \right) (\bar{x}_{ix} - \bar{x}_{ixy}) - \left(\frac{n_{cx}}{n_c} \right) (\bar{x}_{cx} - \bar{x}_{cxy}) \right] \right|.$$

In addition to (c) used in calculating $B1 - B3$ discussed above, the bounds in equations A.10.0–A.10.2 include the following data elements described in section II.C: (d) the means of the baseline measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups (\bar{x}_{ix} , and \bar{x}_{cx}); (e) the means of the

²⁶ In this equation, we ignore an adjustment for the baseline measure. Because the baseline data are imputed, deviations from the MAR assumption can lead to bias in this adjustment. This source of potential bias in the outcome effect size is accounted for separately through the baseline equivalence requirement when data are missing, the technical details of which are discussed in appendix B.

baseline measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups (\bar{x}_{ixy} , and \bar{x}_{cxy}); (f) the standard deviations of the baseline measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data, separately for the intervention and comparison groups, which are used to calculate s_x ²⁷; and (g) the number of subjects with observed baseline data in the analytic sample by condition (n_{cx}).

The formulas for $B1^* - B3^*$ reduce to $B1 - B3$ when there are no missing baseline data.

²⁷ For simplicity, this is referred to using the consistent notation despite the difference in the data used to calculate it.

Appendix B.
Bounding the baseline difference when there are
missing or imputed baseline data

A. Bounding the baseline difference when the outcome is observed for all subjects in the analytic sample

It is not possible to assess baseline equivalence using observed data for the analytic sample in quasi-experimental designs (QEDs) and high-attrition randomized controlled trials (RCTs) that use acceptable approaches to impute baseline data or are missing some baseline data for the analytic sample. However, the What Works Clearinghouse (WWC) will consider the potential bias from baseline differences to be limited if, under different assumptions about whether the data are missing at random (MAR), the standardized baseline difference does not exceed 0.25 standard deviation when the analysis includes an acceptable adjustment for the baseline measure, or 0.05 standard deviation otherwise. This requirement applies only to baseline measures that are required for satisfying the baseline equivalence requirement based on the review protocol.

The WWC uses the same proxy pattern-mixture modeling approach used to address imputed outcome data to estimate the largest possible baseline difference under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

Using the same notation introduced in appendix A, the baseline mean for a sample with missing or imputed baseline data can be modelled using:

$$[B.1] \quad \bar{x}_j = \bar{x}_{jR} + g_j(\rho) \frac{s_x}{s_y} [\bar{y}_j - \bar{y}_{jR}],$$

where \bar{x}_j and \bar{x}_{jR} are the full-sample and complete case baseline means, \bar{y}_j and \bar{y}_{jR} are the full-sample and complete case outcome means, ρ is the correlation between the outcome and the baseline measure, and

$$[B.2] \quad g_j(\rho) = \frac{1}{f_j(\rho)} = \frac{\lambda_j \rho + 1}{\lambda_j + \rho}.$$

The full-sample baseline effect size obtained using an imputation method based on the MAR assumption ($g_c(\rho) = g_i(\rho) = \rho$ when λ_j approaches ∞) can be written as the baseline effect size for the observed sample g_{xR} with an adjustment for the difference between the full-sample and complete case outcome means in the intervention and comparison groups, given by:

$$[B.3] \quad g_{xMAR} = g_{xR} + \frac{\rho}{s_y} ([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}]),$$

where $g_{xR} = \frac{1}{s_x} (\bar{x}_{iR} - \bar{x}_{cR})$. The more general equation for the baseline effect size that allows for deviations from the MAR is:

$$[B.4] \quad g_{xNMAR} = g_{xR} + \frac{1}{s_y} (g_i(\rho) [\bar{y}_i - \bar{y}_{iR}] - g_c(\rho) [\bar{y}_c - \bar{y}_{cR}]).$$

Because $g_j(\rho)$ is bounded between ρ and $1/\rho$, the largest baseline effect size (in absolute value) accounting for deviations from the MAR assumption is given by the maximum of the values given by the following four equations:

$$[B.5.0] \quad C1 = \omega \left| g_{xR} + \frac{\rho}{s_y} ([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}]) \right|$$

$$[B.5.1] \quad C2 = \omega \left| g_{xR} + \frac{1}{\rho s_y} ([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}]) \right|$$

$$[B.5.2] \quad C3 = \omega \left| g_{xR} + \frac{1}{s_y} \rho [\bar{y}_i - \bar{y}_{iR}] - \frac{1}{\rho} [\bar{y}_c - \bar{y}_{cR}] \right|$$

$$[B.5.3] \quad C4 = \omega \left| g_{xR} + \frac{1}{s_y} \frac{1}{\rho} [\bar{y}_i - \bar{y}_{iR}] - \rho [\bar{y}_c - \bar{y}_{cR}] \right|.$$

The first of these, $C1$, is $|g_{xMAR}|$, the estimate of the baseline effect size when MAR holds.

The bounds in equations B.5.0–B.5.3 will be calculated using data reported in studies or obtained from authors. The equations include the following data elements described in section II.C: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups (\bar{y}_i , \bar{y}_c , and the standard deviations are used to calculate the pooled within-group standard deviation s_y); (b) the means of the outcome measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups (\bar{y}_{iR} , \bar{y}_{cR}); (c) the correlation between the baseline and the outcome measures (ρ); and (d) an estimate of the baseline difference based on study data (g_{xR}).

Applying the bounds in equations B.5.0–B.5.3 does not require knowing the baseline effect size using imputed baseline data. Rather, these bounds use the complete case baseline effect size. When the study imputes the baseline data using an acceptable approach and reports the baseline effect size based on imputed data, g_{xI} , a different set of bounds should be used.

Comparing g_{xMAR} and g_{xNMAR} , the bias in the imputed baseline effect size due to deviations from MAR is given by:

$$[B.6] \quad Bias_x = \frac{1}{s_y} \{ (g_i(\rho) - \rho) [\bar{y}_i - \bar{y}_{iR}] - (g_c(\rho) - \rho) [\bar{y}_c - \bar{y}_{cR}] \}.$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size:

$$[B.7.0] \quad D1 = \omega |g_{xI}|$$

$$[B.7.1] \quad D2 = \omega \left| g_{xI} + \frac{1}{s_y} \frac{1-\rho^2}{\rho} [\bar{y}_i - \bar{y}_{iR}] \right|$$

$$[B.7.2] \quad D3 = \omega \left| g_{xI} - \frac{1}{s_y} \frac{1-\rho^2}{\rho} [\bar{y}_c - \bar{y}_{cR}] \right|$$

$$[B.7.3] \quad D4 = \omega \left| g_{xI} + \frac{1}{s_y} \frac{1-\rho^2}{\rho} [(\bar{y}_i - \bar{y}_{iR}) - (\bar{y}_c - \bar{y}_{cR})] \right|.$$

For simplicity, the bounds $C1 - C4$ and $D1 - D4$ were derived based on an imputation model based only on the relationship between the outcome and the baseline measure. If the imputation

model included baseline measures in addition to the outcome, then it is acceptable but not required to replace the outcome means with the average predicted value of the baseline measure. In this case the formula should scale by s_x instead of s_y .

When baseline equivalence is required on multiple baseline measures, the bounds should be calculated separately for each baseline measure, and none may exceed the tolerable thresholds of 0.25 standard deviation when the analysis includes an acceptable adjustment, or 0.05 standard deviation otherwise.

B. Bounding the baseline difference when the outcome measure is imputed for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds $C1 - C4$ or $D1 - D4$. This is because the means of the outcome measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed baseline data.

Similar to the equation for \bar{y}_j in section B of appendix A, the full sample baseline mean for group j can be written as:

$$[B.8] \quad \bar{x}_j = \bar{x}_{jxy} + \left(\frac{n_j - n_{jy}}{n_j} \right) [\bar{x}_{j\sim y} - \bar{x}_{jxy}] + \left(\frac{n_{jy}}{n_j} \right) \left(g_j(\rho) \frac{s_x}{s_y} [\bar{y}_{jy} - \bar{y}_{jxy}] \right),$$

where \bar{x}_{jxy} is the baseline mean for the observations in the complete case analytic sample for group j and is observed at both baseline and for the collection of outcomes, \bar{y}_{jxy} is the outcome mean for the same sample, and \bar{y}_{jy} is the outcome mean for the sample with observed outcome data but possibly missing baseline data.

The baseline effect size obtained using an imputation method based on the MAR assumption ($g_j(\rho) = \rho$) can be written as the difference in the estimated full-sample intervention and comparison group baseline means, given by:

$$[B.9] \quad g_{xMAR} = g_{xR(xy)} + \frac{1}{s_x} \left(\left\{ \left(\frac{n_i - n_{iy}}{n_i} \right) [\bar{x}_{i\sim y} - \bar{x}_{ixy}] + \left(\frac{n_{iy}}{n_i} \right) \frac{\rho s_x}{s_y} [\bar{y}_{iy} - \bar{y}_{ixy}] \right\} - \left\{ \left(\frac{n_c - n_{cy}}{n_c} \right) [\bar{x}_{c\sim y} - \bar{x}_{cxy}] + \left(\frac{n_{cy}}{n_c} \right) \frac{\rho s_x}{s_y} [\bar{y}_{cy} - \bar{y}_{cxy}] \right\} \right).$$

where $g_{xR(xy)} = \frac{1}{s_x} (\bar{x}_{ixy} - \bar{x}_{cxy})$.

The more general formula that allows for deviations from MAR is the following:

$$[B.10] \quad g_{xNMAR} = g_{xR(xy)} + \frac{1}{s_x} \left(\left\{ \left(\frac{n_i - n_{iy}}{n_i} \right) [\bar{x}_{i\sim y} - \bar{x}_{ixy}] + \left(\frac{n_{iy}}{n_i} \right) g_j(\rho) \frac{s_x}{s_y} [\bar{y}_{iy} - \bar{y}_{ixy}] \right\} - \left\{ \left(\frac{n_c - n_{cy}}{n_c} \right) [\bar{x}_{c\sim y} - \bar{x}_{cxy}] + \left(\frac{n_{cy}}{n_c} \right) g_j(\rho) \frac{s_x}{s_y} [\bar{y}_{cy} - \bar{y}_{cxy}] \right\} \right).$$

The largest baseline effect size (in absolute value) accounting for deviations from the MAR assumption is given by the maximum of the values from equations B.11.0–B.11.3:

$$[B.11.0] \quad C1^* = \omega \left| g_{xR(xy)} + \left(\left(\frac{n_i - n_{iy}}{n_i s_x} \right) [\bar{x}_{i \sim y} - \bar{x}_{ixy}] + \rho \left(\frac{n_{iy}}{n_i s_y} \right) [\bar{y}_{iy} - \bar{y}_{ixy}] \right) - \left(\left(\frac{n_c - n_{cy}}{n_c s_x} \right) [\bar{x}_{c \sim y} - \bar{x}_{cxy}] + \rho \left(\frac{n_{cy}}{n_c s_y} \right) [\bar{y}_{cy} - \bar{y}_{cxy}] \right) \right|$$

$$[B.11.1] \quad C2^* = \omega \left| g_{xR(xy)} + \left(\left(\frac{n_i - n_{iy}}{n_i s_x} \right) [\bar{x}_{i \sim y} - \bar{x}_{ixy}] + \frac{1}{\rho} \left(\frac{n_{iy}}{n_i s_y} \right) [\bar{y}_{iy} - \bar{y}_{ixy}] \right) - \left(\left(\frac{n_c - n_{cy}}{n_c s_x} \right) [\bar{x}_{c \sim y} - \bar{x}_{cxy}] + \frac{1}{\rho} \left(\frac{n_{cy}}{n_c s_y} \right) [\bar{y}_{cy} - \bar{y}_{cxy}] \right) \right|$$

$$[B.11.2] \quad C3^* = \omega \left| g_{xR(xy)} + \left(\left(\frac{n_i - n_{iy}}{n_i s_x} \right) [\bar{x}_{i \sim y} - \bar{x}_{ixy}] + \rho \left(\frac{n_{iy}}{n_i s_y} \right) [\bar{y}_{iy} - \bar{y}_{ixy}] \right) - \left(\left(\frac{n_c - n_{cy}}{n_c s_x} \right) [\bar{x}_{c \sim y} - \bar{x}_{cxy}] + \frac{1}{\rho} \left(\frac{n_{cy}}{n_c s_y} \right) [\bar{y}_{cy} - \bar{y}_{cxy}] \right) \right|$$

$$[B.11.3] \quad C4^* = \omega \left| g_{xR(xy)} + \left(\left(\frac{n_i - n_{iy}}{n_i s_x} \right) [\bar{x}_{i \sim y} - \bar{x}_{ixy}] + \frac{1}{\rho} \left(\frac{n_{iy}}{n_i s_y} \right) [\bar{y}_{iy} - \bar{y}_{ixy}] \right) - \left(\left(\frac{n_c - n_{cy}}{n_c s_x} \right) [\bar{x}_{c \sim y} - \bar{x}_{cxy}] + \rho \left(\frac{n_{cy}}{n_c s_y} \right) [\bar{y}_{cy} - \bar{y}_{cxy}] \right) \right|.$$

In addition to (c) and (d) used in calculating $C1^*–C4^*$, the bounds in equations B.11.0–B.11.3 include the following data elements described in section II.C: (e) the means of the outcome measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups (\bar{y}_{iy} , and \bar{y}_{cy}); (f) the means of the outcome measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups (\bar{y}_{ixy} , and \bar{y}_{cxy}); (g) the standard deviations of the outcome measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data, which are used to calculate s_x ²⁸; and (h) the number of subjects with observed outcome data in the analytic sample by condition (n_i , and n_c).

Applying the bounds $C1^*–C4^*$ does not require knowing the baseline effect size using imputed baseline data. Rather, these bounds use the complete case baseline effect size. When the study imputes the baseline data using an acceptable approach and reports the baseline effect size based on imputed data, g_{xI} , a different set of bounds should be used.

Comparing g_{xMAR} and g_{xNMAR} , the bias in the imputed baseline effect size due to deviations from MAR is given by:

$$[B.12] \quad Bias_x = \frac{1}{s_y} \left\{ \left(\frac{n_{iy}}{n_i} \right) (g_i(\rho) - \rho) [\bar{y}_{iy} - \bar{y}_{ixy}] - \left(\frac{n_{cy}}{n_c} \right) (g_c(\rho) - \rho) [\bar{y}_{cy} - \bar{y}_{cxy}] \right\}.$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size $D1^*–D4^*$:

²⁸ For simplicity, this is referred to using the consistent notation despite the difference in the data used to calculate it.

$$[B.13.0] \quad D1^* = \omega |g_{xI}|$$

$$[B.13.1] \quad D2^* = \omega \left| g_{xI} + \frac{1}{s_y} \left(\frac{n_{iy}}{n_i} \right) \frac{1-\rho^2}{\rho} [\bar{y}_{iy} - \bar{y}_{ixy}] \right|$$

$$[B.13.2] \quad D3^* = \omega \left| g_{xI} - \frac{1}{s_y} \left(\frac{n_{cy}}{n_c} \right) \frac{1-\rho^2}{\rho} [\bar{y}_{cy} - \bar{y}_{cxy}] \right|$$

$$[B.13.3] \quad D4^* = \omega \left| g_{xI} + \frac{1}{s_y} \frac{1-\rho^2}{\rho} \left[\left(\frac{n_{iy}}{n_i} \right) (\bar{y}_{iy} - \bar{y}_{ixy}) - \left(\frac{n_{cy}}{n_c} \right) (\bar{y}_{cy} - \bar{y}_{cxy}) \right] \right|.$$

The formulas for $C1^* - C4^*$ and $D1^* - D4^*$ reduce to $C1 - C4$ and $D1 - D4$ when there are no missing outcome data.

Appendix C.
Additional detail for reviews of studies
that present complier average causal effects estimates

A. Conceptual background for rating complier average causal effects estimates when attrition is low

1. Criterion 1: No clear violations of the exclusion restriction

Under the exclusion restriction, the only channel through which assignment to the intervention or comparison groups can influence outcomes is by affecting take-up of the intervention being studied (Angrist et al., 1996). The exclusion restriction implies that always-takers in the intervention and comparison groups should not differ in outcomes because their assignment status did not influence their take-up status; likewise, never-takers in the intervention and comparison groups should not differ in outcomes. When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may be attributable to other factors differing between the intervention and comparison groups.

The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels. However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.

Existing What Works Clearinghouse (WWC) standards that prohibit “confounding factors”—factors that differ completely between the assigned groups—already rule out many violations of the exclusion restriction. For example, if groups differ in their eligibility for interventions other than the intervention being studied, then the implied violation of the exclusion restriction is also a confounding factor that, under current WWC group design standards, would cause a study to be rated *Does Not Meet WWC Group Design Standards*.

One scenario that does not represent a confounding factor in intent-to-treat (ITT) studies would be a violation of the exclusion restriction. The exclusion restriction would be violated if take-up was defined inconsistently between the assigned intervention group and assigned comparison group. For example, suppose that take-up in the assigned intervention group was defined as enrolling in the intervention being studied, such as an intensive afterschool program, whereas take-up in the assigned comparison group was defined as enrolling in the specified intervention *or* “similar” interventions, such as attending any program after school. In this case, differences in outcomes between assigned groups might not be attributable solely to differences in rates of take-up as defined by the study because the two take-up rates measure different concepts.

Another violation of the exclusion restriction that does not necessarily stem from a confounding factor is the scenario in which assignment to the intervention group changes the behavior of subjects even if they do not take up the intervention itself. For example, in an experiment to test the effectiveness of requiring unemployed workers to receive job-search and training services, assignment to the intervention group might motivate subjects to search for a job to avoid having to participate in the intervention services. In this case, the intervention assignment might have effects on outcomes through channels other than the take-up rate.

Judgment is required to determine whether a potential unintended channel for group status to influence outcomes is important enough to undermine the internal validity of a complier average causal effects (CACE) estimate. Under this guidance, the WWC's lead methodologist for a review has the responsibility to make this judgment.

2. Criterion 2: Sufficient instrument strength

The condition of sufficient instrument strength requires that the group assignment indicators—that is, the instrumental variables—collectively serve as strong predictors of take-up, the endogenous independent variable. As discussed next, this condition is necessary for conventional statistical tests based on two-stage least squares (2SLS) estimators to have low type I (false positive) error rates.

The need for sufficient instrument strength stems from the statistical properties of 2SLS estimators. An extensive statistical literature has demonstrated that, in finite samples, 2SLS estimators of CACE impacts include part of the bias of ordinary least-squares estimates (Basmann, 1974; Bloom, Zhu, & Unlu, 2010; Bound, Jaeger, & Baker, 1995; Buse, 1992; Nelson & Startz, 1990; Richardson, 1968; Sawa, 1969).²⁹ Moreover, in finite samples, 2SLS estimators do not have a normal distribution—the distribution typically used to construct confidence intervals. For these reasons, conventional statistical tests—such as *t* tests and *F* tests—based on 2SLS estimators in finite samples have actual type I error rates that generally are higher than the assumed type I error rates (Stock & Yogo, 2005). For instance, a *t* test conducted at an assumed 5 percent significance level will have an actual type I error rate exceeding 5 percent.

The bias issue with 2SLS estimators shrinks as the instruments become stronger predictors of the endogenous independent variable. An instrument is considered a stronger predictor of an endogenous independent variable if the association between the instrument and endogenous independent variable is larger or the association is more precisely estimated. In the context of estimating CACE effects, group status is a stronger instrument when group take-up rates differ more and when sample sizes are larger.

Instruments also must be strong enough for statistical tests of 2SLS estimators to have “acceptably” low type I error rates. As instruments become stronger, the probability distributions of 2SLS estimators converge to normal distributions centered on the true CACE impact. Type I error rates follow suit and converge to their assumed levels. We put “acceptably” in quotes because defining what is acceptable requires its own standard, which is explained next.

Selecting the maximum tolerable type I error rate is the first step in establishing a criterion for sufficient instrument strength. WWC standards do not provide a precedent for acceptable rates of type I error but do provide a precedent for acceptable levels of bias in impact estimates, which is

²⁹ As discussed by Bloom et al. (2010), the finite-sample bias of instrumental variable estimators originates from sampling error. Due to finite samples, random assignment will produce intervention and comparison groups that, by chance, are not fully identical on the characteristics of group members. Some of these unobserved characteristics exert influences on *both* take-up and outcomes. For illustrative purposes, suppose take-up and outcomes are positively correlated due to these unobserved influences. When sampling error leads to greater (or smaller) differences in take-up between the intervention and comparison groups, greater (or smaller) differences in outcomes arise. Although both types of differences result from random imbalances, the differences are systematically related, creating a spurious association between take-up and outcomes.

0.05 standard deviation. We use this precedent to set acceptable type I error rates. In the next section, we present a statistical framework that links type I error rates to estimation bias. Using this framework, for a *t*-test whose assumed type I error rate is 0.05, ensuring a bias of less than 0.05 standard deviation implies actual type I error rates of less than 0.10. Thus, the guidelines for instrument strength specified here are based on an upper limit of 0.10 for the type I error rate.

B. Linking complier average causal effects estimation bias with type I error rates

In this section, we provide a statistical framework for deriving the relationship between the bias of an impact estimator and the estimator's type I error rate. We focus on a conventional *t* test. In this framework, setting a maximum tolerable bias—for which there is precedent in WWC standards—implies setting a maximum tolerable type I error rate.

Consider a situation in which the true impact of an intervention β_1 is zero. A biased estimator of this impact $\hat{\beta}_1^{biased}$ will have a distribution centered on a value different from zero. Larger bias increases type I error; as the distribution of the estimator lies further away from zero, there is a greater likelihood of incorrectly rejecting the hypothesis of a zero impact, assuming correct variances are estimated.

To derive the relationship between bias and type I error rates, we cannot use the distribution of the 2SLS estimator because its distribution has neither an expected value, when only one instrument is employed, nor a familiar distribution in finite samples (Stock, Wright, & Yogo, 2002). Instead, we consider a generic estimator expressed in effect size units $\hat{\beta}_1^{biased}$. It is distributed normally with expected value equal to $b > 0$ standard deviations when the true impact is zero. The probability of a type I error using a 5 percent significance test is

$$\begin{aligned}
 \text{[C.1.0]} \quad \text{Type I error rate} &= \Pr\left(\frac{\hat{\beta}_1^{biased}}{SE(\hat{\beta}_1^{biased})} > z_{0.975}\right) + \Pr\left(\frac{\hat{\beta}_1^{biased}}{SE(\hat{\beta}_1^{biased})} < z_{0.025}\right) \\
 &= \Pr\left(\frac{\hat{\beta}_1^{biased} - b}{SE(\hat{\beta}_1^{biased})} > \left(z_{0.975} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right)\right) + \Pr\left(\frac{\hat{\beta}_1^{biased} - b}{SE(\hat{\beta}_1^{biased})} < \left(z_{0.025} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right)\right) \\
 &= 1 - \Phi\left(z_{0.975} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right) + \Phi\left(z_{0.025} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right),
 \end{aligned}$$

where $SE(\bullet)$ denotes the standard error of an estimator, z_q is the q th quantile of the standard normal distribution, and $\Phi(\bullet)$ is the cumulative distribution function of the standard normal distribution.

Equation C.1.0 provides the relationship between the type I error rate and bias as long as the standard error of the biased estimator is known. Therefore, to specify this relationship fully, we must pick a value for the standard error. The standard error can vary depending on sample size, covariates, degree of clustering, and other factors. Picking a standard error essentially entails choosing a “benchmark” level of precision to complete the specification of equation C.1.0.

As the benchmark, we assume a level of precision corresponding to a study for which the minimum detectable effect size is 0.25 standard deviation. A value for minimum detectable effect size, in turn, directly implies a value for the standard error. Specifically, the minimum effect size that can be detected using a two-tailed test at a 5 percent significance level with

80 percent power can be expressed as a function of the standard error (SE), as follows (see Bloom, 2004):

$$[C.2] \quad MDES = [\Phi^{-1}(1 - 0.05/2) + \Phi^{-1}(0.8)] \times SE = 2.802 \times SE.$$

Using equation C.2, a study designed to have a minimum detectable effect size of 0.25 is expected to have a standard error of 0.09 standard deviation ($= 0.25 / 2.802$).

By substituting the benchmark standard error, 0.09 standard deviation, for $SE(\hat{\beta}_1^{biased})$ in equation C.1.0, we completely specify the relationship between the type I error rate and the amount of bias. Equation C.1.0 becomes

$$[C.1.1] \quad \text{Type I error rate} = 1 - \Phi(z_{0.975} - b/0.09) + \Phi(z_{0.025} - b/0.09).$$

The final step is to substitute into equation C.1.1 a value for b that represents the maximum tolerable bias. As discussed earlier, the maximum value for b that is acceptable to the WWC is 0.05 standard deviation. Setting $b = 0.05$ in equation C.1.1, we obtain a maximum tolerable type I error rate equal to

$$\text{Maximum tolerable type I error rate} = 1 - \Phi(z_{0.975} - 0.05/0.09) + \Phi(z_{0.025} - 0.05/0.09) = 0.086.$$

The maximum tolerable type I error rate then determines the minimum required first-stage F statistic for sufficient instrument strength. For a given number of instruments, Stock and Yogo (2005) calculated several different values for the minimum required first-stage F statistic, depending on whether the maximum tolerable type I error rate is 0.10, 0.15, 0.20, or 0.25. For setting the WWC standard, our preceding calculations yield a maximum tolerable type I error rate of 0.086, which we round to 0.10, the closest value addressed by Stock and Yogo (2005). We then use this value to produce values for the minimum required first-stage F statistic based on Stock and Yogo's (2005) calculations.

C. Calculating attrition and baseline differences when there are three or more groups to which each sample member could be randomly assigned in a complier average causal effects analysis

1. Calculating attrition

Section II.D of this *Standards Handbook* provides formulas for calculating the overall and differential attrition rate for compliers when there are two assigned groups (the intervention group and comparison group). In appendix C.C.1, we consider the scenario in which there are three or more groups to which each sample member could be randomly assigned, for instance, a group that is ineligible for the intervention, a group that has low priority for the intervention, and a group that has high priority for the intervention. Even though there are multiple assigned groups, there is still only a single intervention being studied, so there is still only a single measure of take-up—a binary variable for taking up any portion of the intervention.

First, we order the assigned groups with the index $k = 0, 1, 2, \dots, K$ from lowest to highest take-up rate. We also make a monotonicity assumption (Imbens & Angrist, 1994): Any sample

member who would take up the intervention if assigned to group k would also take up the intervention if assigned to a group ordered after k . For each comparison between group $(k - 1)$ and group k , compliers are defined as those who would take up the intervention if assigned to group k but not if assigned to group $(k - 1)$. The 2SLS estimator of the CACE is a weighted average of complier impacts across these comparisons, with weights given by Imbens and Angrist (1994). Therefore, our method for calculating attrition follows the same approach: We calculate attrition (both overall and differential) for each comparison between consecutively ordered groups, and then take a weighted average across those comparisons, using the same weights as those in the 2SLS estimator.

Specifically, let $\widehat{\Delta}_{k,k-1}^{complier}$ be the differential attrition rate for compliers pertaining to the comparison between groups $(k - 1)$ and k , based on applying equation II.1. The final differential attrition rate for all compliers $\widehat{\Delta}_{final}^{complier}$ is calculated as

$$[C.3] \quad \widehat{\Delta}_{final}^{complier} = \frac{\sum_{k=1}^K \lambda_k \widehat{\Delta}_{k,k-1}^{complier}}{\sum_{k=1}^K \lambda_k},$$

where λ_k is the weight on the comparison between groups $(k - 1)$ and k . Imbens and Angrist (1994) derived the weight to be

$$[C.4] \quad \lambda_k = (\bar{D}_{k,ran} - \bar{D}_{k-1,ran}) \sum_{l=k}^K \frac{N_l}{N} (\bar{D}_{l,ran} - \bar{D}_{ran}),$$

where $\bar{D}_{k,ran}$ is the take-up rate for sample members assigned to group k , \bar{D}_{ran} is the take-up rate in the entire randomly assigned sample, N_k is the number of sample members assigned to group k , and N is the total number of sample members in the entire randomly assigned sample.

For calculating overall attrition, the same weights are used to take a weighted average of the overall complier attrition rates across all comparisons.

2. Calculating baseline differences

The final calculation of a baseline difference on a characteristic specified in the protocol follows a similar approach as that used for calculating attrition. For each comparison between groups $(k - 1)$ and k , we use equation II.6 to calculate the baseline difference for compliers in the analytic sample. We then take a weighted average of those baseline differences. The weight on each comparison is again specified by equation C.4, except that all sample sizes and take-up rates are calculated from the analytic sample, not the original randomly assigned sample.

References

Allison, P. D. (2002). *Missing data* (Paper No. 136). Thousand Oaks, CA: Sage University.

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1), 46–63.

Andridge, R. R., & Little, R. J. A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27(2), 153–180.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–472.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work. *International Journal of Methods in Psychiatric Research*, 20(1), 40–49.

Basman, R. (1974). Exact finite sample distributions for some econometric estimators and test statistics: A survey and appraisal. In M. Intriligator & D. Kendrick (Eds.), *Frontiers of quantitative economics*, Vol. 2 (pp. 209–288). Amsterdam: North Holland Publishing Co.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency*, 83, 460–472.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8(2), 225–246.

Bloom, H. (2004). *Randomizing groups to evaluate place-based programs*. New York, NY: MDRC.

Bloom, H., Zhu, P., & Unlu, F. (2010). *Finite sample bias from instrumental variables analysis in randomized trials*. New York, NY: MDRC.

Bound, J., Jaeger, D., & Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430), 443–450.

Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica*, 60, 173–180.

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, 82(6), 2295–2326.

Cragg, J., & Donald, S. (1993). Testing identifiability and specification in instrumental variables models. *Econometric Theory*, 9(2), 222–240.

Fier, D., Lemieux, T., & Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business and Economic Statistics*, 34(2), 185–196.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In S. N. Haynes & E. M. Hieby (Eds.), *Comprehensive handbook of psychological assessment. Vol. 3: Behavioral assessment* (pp. 108–127). New York, NY: Wiley.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6(2), 107–128.

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Expanding analysis and use of single-case research. *Education and Treatment of Children*, 35, 269–290.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–475.

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3), 933–959.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2), 615–635.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25, 305–327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, 64(4), 555–574.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433), 222–230.

Lee, D., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2), 655–674.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., et al. (2012). The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine*, 367(14), 1355–1360.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

McCrory, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698–714.

Nelson, C., & Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, 58(4), 967–976.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Reardon, S., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, 5(1), 83–104.

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association*, 63(324), 1214–1226.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Sanderson, E., & Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2), 212–221.

Sawa, T. (1969). The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association*, 64(325), 923–937.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8(1), 3–15.

Schochet, P. Z. (2008). *The late pretest problem in randomized control trials of education interventions* (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 64(3), 557–586.

Stock, J., Wright, J., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20(4), 518–529.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In J. Stock & D. W. K. Andrews (Eds.), *Identification and inference for econometric models: Essays in Honor of Thomas J. Rothenberg* (pp. 80–108). Cambridge, MA: Cambridge University Press.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Wolf, A., Price, C., Miller, H., & Boulay, B. (2017). *Establishing baseline equivalence: A practical guide for evaluators*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, 38(2), 107–141.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

January 2020

This report was prepared for the Institute of Education Sciences (IES) under Contract 91990018C0019 by the American Institutes for Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This document is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

What Works Clearinghouse. (2020). What Works Clearinghouse Standards Handbook, Version 4.1. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at <https://ies.ed.gov/ncee/wwc/handbooks>.